

**Towards understanding mastrevirus dynamics and the
use of viral metagenomic approaches to identify novel
gemini-like circular DNA viruses**



A thesis
submitted in fulfilment
of the requirements for the Degree
of

Doctor of Philosophy

at the
University of Canterbury
New Zealand

Simona Kraberger

2014

Contents

Acknowledgements.....	III
Abstract.....	IV
Co-authorship forms.....	VII
Chapter 1 Literature review.....	1
Chapter 2 Australian monocot-infecting mastrevirus diversity rivals that in Africa.....	75
Chapter 3 Molecular diversity of monocot-infecting mastreviruses in Africa.....	114
Chapter 4 Evidence that dicot-infecting mastreviruses are particularly prone to inter-species recombination and have likely been circulating in Australia for longer than in Africa and the Middle East.....	151
Chapter 5 Identification of an Australian-like dicot-infecting mastrevirus in Pakistan.....	185
Chapter 6 Molecular diversity of <i>Chickpea chlorotic dwarf virus</i> in Sudan: high rates of intra-species recombination - a driving force in the emergence of new strains.....	201
Chapter 7 Identification of novel circular DNA viruses associated with <i>Poaceae</i> species in New Zealand.....	239
Chapter 8 Characterisation of a diverse range of Rep-encoding ssDNA viruses recovered from a sewage treatment oxidation pond.....	260
Chapter 9 Discussion and future directions.....	304

Acknowledgments

I would like to thank the following people for their support and encouragement throughout this thesis journey, without them this thesis would not have been possible.

Arvind Varsani for taking me on as one of his first PhD students at the University of Canterbury and being the best supervisor anyone could ask for. His enthusiasm for science and life in general is contagious and I cannot express enough how truly grateful I am for his ongoing patience, wealth of knowledge, humour and support. Arvind has encouraged me to constantly ask questions, learn new skills, and has given me the opportunity to travel and partake in a broad range of research projects in addition to my PhD thesis research, for which I will always be thankful.

My co-supervisors; **Darren Martin** for his guidance, expertise and invaluable contributions to this thesis and **Dave Collings** for his help and advice whenever I needed it.

All the wonderful scientists around the world who provided me with plant samples, without these contributions this research would not have been possible.

My amazing parents **Angelika** and **Robert**, and the rest of my family **Georgina**, **Juliet**, **Oskar**, **Johann**, and **Steffan**, for always supporting and believing in me.

Daniel words cannot express how thankful I am for your continuous encouragement and support.

My fellow lab-mates in the Varsani Lab and beyond, with a special thanks to **Anisha**, **Daisy**, **Dorien**, **Ryan**, **Jacqui**, **Katherine**, **Laurel** and **Dean**. All have been hugely supportive in their own unique ways, reminding me to not take life too seriously and providing a constant supply of chocolate and laughs.

All my true friends near and far (you know who you are) thank you for nodding and smiling when I talk about my research and never losing faith in me even when I was temporarily absent from life in my thesis bubble.

Abstract

Mastreviruses (family *Geminiviridae*) are plant-infecting viruses with circular single-stranded (ss) DNA genomes (~2.7kb). The genus *Mastrevirus* is comprised of thirty-two species which are transmitted by leafhoppers belonging to the genus *Cicadulina*. Mastreviruses are widely distributed and have been found in the Middle East, Europe, Asia, Australia, Africa and surrounding islands. Only one species, dragonfly-associated mastrevirus has so far been identified in the Americas, isolated from a dragonfly in Puerto Rico. Species can be group based on the host(s) they infect, those which infect monocotyledonous (monocot) plants and those which infect dicotyledonous (dicot) plants. In recent years many new mastrevirus species have been discovered. Several of these new discoveries can largely been attributed to the development of new molecular tools. The current state of sequencing platforms has made it affordable and easier to characterise mastreviruses at a genome level thus allowing scientists to delve deeper into understanding the dynamics of mastreviruses. A few mastrevirus species have been identified as important agricultural pathogens and as a result have been the focus of much of the mastrevirus research. *Maize streak virus*, strain A (MSV-A) has been the most extensively studied due to the devastating impact it has on maize production in Africa. Studies have shown that MSV-A likely emerged as a pathogen of maize less than 250 years following introduction of maize in Africa by early European settlers. There is compelling evidence to suggest that MSV-A is likely the result of recombination events between wild grass adapted MSV strains. It therefore is equally important to monitor viruses infecting non-cultivated plants in order to gain a greater understanding of the epidemiological dynamics of mastreviruses, which in turn is essential for implementing disease management strategies.

The objective of the research undertaken as part of this PhD thesis was to investigate global mastrevirus dynamics focusing on diversity, host and geographic ranges, mechanisms of evolution, phylogeography and possible origins of these viruses. In addition to this a viral metagenomic approach was used in order to identify novel mastreviruses or mastrevirus-like present in New Zealand.

The dynamics of the monocot-infecting mastreviruses are investigated in Chapter Two and Three. The work described in these two chapters focus mainly on mastreviruses which infect non-cultivated grasses in Africa and Australia, a total of 161 full mastrevirus genomes were recovered collectively in the two studies. Chapter Two reveals a high level of mastrevirus diversity present in Australia with the discovery of four new species and several new strains of previously characterised species. An extensive sampling effort in Africa undertaken in Chapter Three reveals a broader host range and geographic distribution of the African monocot-infecting mastreviruses than previously documented. Mosaic patterns of recombination are evident among both the Australian and African monocot-infecting mastreviruses.

In Chapters Four, Five and Six a comprehensive investigation was undertaken focusing on the dicot-infecting mastreviruses. The study undertaken in Chapter Four entailed the recovery of 49 full mastrevirus genomes from Australia, the Middle East, Africa, Turkey and the Indian Subcontinent to investigate the diversity of dicot-infecting mastreviruses from a global context. Analyses revealed a high degree of CpCDV strain diversity and extended the known geographic range of CpCDV. For the first time phylogeographic analysis was able to investigate the origins of the dicot-infecting mastreviruses. Results revealed the likely origin of the most recent common ancestor (MRCA) of these viruses is likely closer to Australia than anywhere else that dicot-infecting mastreviruses have been sampled and illuminated a supported series of historical movements following the emergence of the MRCA. In Chapter Five two novel mastreviruses Australian-like mastreviruses were isolated from chickpea material from Pakistan. A comprehensive analysis of CpCDV isolates in the major pulse growing regions of Sudan in Chapter Six reveals that this region harbours a high degree of strain diversity. Complex patterns of intra-species recombination indicate these strains are evidently circulating in these regions and infecting the same hosts, driving the emergence of new CpCDV strains.

Collectively the results discussed in Chapters Two through Six extended the current knowledge of mastrevirus diversity. The natural host range of many mastreviruses has

proven to be more extensive than previously documented, with many species having overlapping host ranges and hence these hosts could be acting as ‘mixing vessels’ enabling inter-species recombination. Patterns of recombination and selection were observed in both the monocot-infecting and the dicot-infecting mastreviruses further elucidating the mechanisms these viruses employ to evolve rapidly. Extensive sampling in a wide range of geographic regions provides insights into the true geographic range of species such as MSV and CpCDV.

Given that mastreviruses have been able to move globally and Australia has been identified as a major mastrevirus diversity hotspot it is conceivable that mastreviruses are also present in New Zealand. In Chapter Seven and Eight this is explored by using a viral metagenomic approach to investigate the ssDNA viral populations associated with wild grasses and sewage material in New Zealand. Although no mastreviruses were recovered, this endeavour resulted in the discovery of more than 50 novel circular Rep-encoding ssDNA (CRESS DNA) viruses associated with non-cultivated grasses and treated sewage material, many of which are similar to mastreviruses and other geminiviruses. These discoveries expand current knowledge on the diversity of ssDNA viruses present in New Zealand and further highlight this viral metagenomic approach as an effective method for ssDNA virus discovery.

Overall the results discussed in this thesis provide insights into mastrevirus diversity and dynamics as well as revealing a wealth of novel CRESS DNA viruses, some of which share similarities to geminiviruses.

Deputy Vice-Chancellor's Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 2

Kraberger, S., Thomas, J.E., Geering, A.D.W., Dayaram, A., Stainton, D., Hadfield, J., Walters, M., Parmenter, K.S., van Brunschot, S., Collings, D.A., Martin, D.P. and Varsani, A. (2012) Australian monocot-infecting mastrevirus diversity rivals that in Africa. *Virus Research* 169(1), 127-136.

Please detail the nature and extent (%) of contribution by the candidate:

The samples were collected by were collected as a team effort by Simona Kraberger, Andrew Geering, John Thomas, Daisy Stainton, Anisha Dayaram, James Hadfield, Matt Walters, Kathy Parmenter, Sharon von Brunschot and David Collings over the period the Varsani lab relocated to Brisbane (Australia) due to lab shut down as a consequence of earthquakes.

The all molecular work was undertaken by Simona Kraberger. Martin and Varsani double checked all the bioinformatics analysis.

Simona Kraberger wrote the manuscript and the rest of the authors provided comments / feedback to improve it.

Contribution by Simona Kraberger: 90%

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Arvind Varsani* Signature:



Date: 24th Oct 2014

Deputy Vice-Chancellor's Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 4

Krabberger, S., Harkins, G.W., Kumari, S.G., Thomas, J.E., Schwinghamer, M.W., Sharman, M., Collings, D.A., Briddon, R.W., Martin, D.P. and Varsani, A. (2013) Evidence that dicot-infecting mastreviruses are particularly prone to inter-species recombination and have likely been circulating in Australia for longer than in Africa and the Middle East. *Virology* 444(1–2), 282-291.

Please detail the nature and extent (%) of contribution by the candidate:

The samples used in this study were collected by Saafa Kumari, Mark Schwinghamer, Murray Sharman and Rob Briddon.

The all molecular work was undertaken by Simona Kraberger. Simona Kraberger worked with Gordon Harkins to undertake phylogeography analysis described in this chapter. Darren Martin and Arvind Varsani double checked all the bioinformatics analysis.

Simona Kraberger wrote the manuscript and the rest of the authors provided comments / feedback to improve it.

Contribution by Simona Kraberger: 85%

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Arvind Varsani* Signature:



Date: 24th October 2014

Deputy Vice-Chancellor's Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 5

Kraberger, S., Mumtaz, H., Claverie, S., Martin, D. P., Briddon, R. W., Varsani, A. (2014)
Identification of an Australian-like dicot-infecting mastrevirus in Pakistan. *Archives of Virology*
160, 825-830.

Please detail the nature and extent (%) of contribution by the candidate:

The samples used in this chapter were collected by Huma Mumtaz and Rob Briddon.

The all molecular work was undertaken by Simona Kraberger. Darren Martin and Arvind Varsani double checked all the bioinformatics analysis.

Simona Kraberger wrote the manuscript and the rest of the authors provided comments / feedback to improve it.

Contribution by Simona Kraberger: 90%

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Arvind Varsani* Signature:



Date: *24th Oct 2014*

Deputy Vice-Chancellor's Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 6

Kraberger, S., Kumari, S., Hamed, A. A., Gronenborn, B., Thomas, J. E., Sharman, M., Harkins, G. W., Muhire, B. M., Martin, D. P., Varsani, A. (2014) Molecular diversity of *Chickpea chlorotic dwarf virus* in Sudan: high rates of intra-species recombination a driving force in the emergence of new strains. *Infection, Genetics and Evolution* 29, 203-215.

Please detail the nature and extent (%) of contribution by the candidate:

The samples used in this study were collected by Saafa Kumari, Abdelmajid Hamed and Bruno Gronenborn. John Thomas and Murray Sharman undertook pre-screening of the samples using serology before they were forwarded to the Varsani lab.

The all molecular work was undertaken by Simona Kraberger. Darren Martin and Arvind Varsani double checked all the bioinformatics analysis.

Simona Kraberger wrote the manuscript and the rest of the authors provided comments / feedback to improve it.

Contribution by Simona Kraberger: 90%

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Arvind Varsani* Signature:



Date: 24th October 2014

Deputy Vice-Chancellor's Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 7

Kraberger, S., Farkas, K., Bernardo, P., Booker, C., Argüello-Astorga, G. R., Mesléard, F., Martin, D. P., Varsani, A. (2015) Identification of novel Bromus- and Trifolium-associated circular DNA viruses. Archives of Virology. DOI 10.1007/s00705-015-2358-6

Please detail the nature and extent (%) of contribution by the candidate:

The samples used in this study were collected by Simona Kraberger, Arvind Varsani and Cameron Booker.

The all molecular work was undertaken by Simona Kraberger. Gerardo Argüello-Astorga worked with Simona Kraberger to identify putative iterons in the sequences. Simona Kraberger worked with Arvind Varsani to analyse the next-generation sequence data. Darren Martin and Arvind Varsani double checked all the bioinformatics analysis.

Simona Kraberger wrote the manuscript and the rest of the authors provided comments / feedback to improve it.

Contribution by Simona Kraberger: 95%

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Arvind Varsani* Signature:



Date: 24th October 2014

Deputy Vice-Chancellor's Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 8

Krabberger, S., Argüello-Astorga, G. R., Greenfield, L. G., Galilee, C., Law, D., Martin, D. P., Varsani, A., (2015) Characterisation of a diverse range of Rep-encoding ssDNA viruses recovered from a sewage treatment oxidation pond. *Infection, Genetics and Evolution* 31, 73-86.

Please detail the nature and extent (%) of contribution by the candidate:

The sample was collected by Donald law with the help of Craig Galilee (restricted access to the site).

The all molecular work was undertaken by Simona Kraberger. Simona Kraberger worked with Arvind Varsani to analyse the next-generation sequence data. Simona Kraberger worked with Gerardo Argüello-Astorga to identify iterons and putative nonanucleotide motifs in the viral genomes required for initiation of singles-stranded DNA virus genome replication in the sequences. Darren Martin and Arvind Varsani double checked all the bioinformatics analysis.

Simona Kraberger wrote the manuscript and the rest of the authors provided comments / feedback to improve it.

Contribution by Simona Kraberger: 90%

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Arvind Varsani* Signature:



Date: 24th Oct 2014

Chapter 1

Literature review

Contents

1.1	Introduction.....	3
1.2	Geminiviruses	4
1.2.1	The genus <i>Begomovirus</i>	9
1.2.2	The genus <i>Curtovirus</i>	10
1.2.3	The genus <i>Topocuvirus</i>	10
1.2.4	The genus <i>Becurtovirus</i>	11
1.2.5	The genus <i>Eragrovirus</i>	11
1.2.6	The genus <i>Turncurtovirus</i>	12
1.2.7	The genus <i>Mastrevirus</i>	12
1.2.8	Unclassified highly divergent geminivirus.....	13
1.2.9	Replication of geminiviruses	16
1.2.10	Theories behind the evolutionary origin of geminiviruses.....	19
1.2.11	Geminivirus evolution	20
1.2.11.1	Genetic drift.....	20
1.2.11.2	Recombination and reassortment	22
1.2.12	Genome secondary structure	23
1.3	Mastreviruses.....	24
1.3.1	Genomes of mastreviruses	24
1.3.2	Intergenic regions.....	24
1.3.3	Movement protein (V2)	27
1.3.4	Capsid protein (V1).....	27
1.3.5	Replication-associated protein and RepA (C1 and C1:C2)	29
1.3.6	Monocot-infecting mastreviruses.....	33
1.3.6.1	African streak mastreviruses.....	33
1.3.6.2	Australian striate mosaic mastreviruses.....	34
1.3.6.3	Japan-Pacific mastreviruses	36

1.3.6.4	Eurasian mastreviruses.....	36
1.3.7	Dicot-infecting mastreviruses	36
1.4	Mastrevirus detection methods	41
1.4.1	Serology	41
1.4.2	Polymerase chain reaction	41
1.4.3	Rolling circle amplification	41
1.5	Next-Generation sequencing (NGS).....	42
1.5.1	Roche/454's GS FLX system.....	42
1.5.2	Illumina/Solexa's GA HiSeq system	43
1.5.3	Biosystems/SOLiD system	44
1.6	NGS approaches for the discovery of novel geminiviruses and other Rep encoding ssDNA viruses	44
1.7	Aims and rational of this study	47
1.8	References.....	54

1.1 Introduction

Viruses infect organisms from all kingdoms of life and are found in all ecosystems, from the thermal vents at the bottom of the ocean (Geslin *et al.*, 2003; Ortmann & Suttle, 2005; Williamson *et al.*, 2008) to the Dry valleys of Antarctica (Kepner Jr *et al.*, 1998; Laybourn Parry *et al.*, 2001; Swanson *et al.*, 2012; Takacs & Priscu, 1998; Zawar-Reza *et al.*, 2014), the Sahara Desert (Prigent *et al.*, 2005) and even hyper saline environments (Atanasova *et al.*, 2012; Krupovic *et al.*, 2011; Roine *et al.*, 2010). The diverse nature of viruses and their abundance is apparent from the daily discoveries of novel viruses, however, it is evident when considering the quantity of novel viruses being discovered through viral metagenomic studies (Labonté & Suttle, 2013; Ng *et al.*, 2011a; Ng *et al.*, 2012; Rosario *et al.*, 2009a; Roux *et al.*, 2012; Whon *et al.*, 2012) that the true extent of the global viral diversity is immensely underestimated with possibly less than 1% of all viruses on earth being catalogued and/or characterised. The classification of viruses, proposed by David Baltimore, broadly categorises viruses into seven groups based on their genetic make-up, their approach to generate mRNA and replication strategy (Brown *et al.*, 2011). Following this the international committee for virus taxonomy (ICTV) was established to implement taxonomic guidelines for the classification of viruses. A grouping which in recent years has seen increased activity in terms of discovery of novel viruses is the single-stranded DNA (ssDNA) viruses. Advances in molecular techniques and sequencing technologies have enable scientists to unravel some of the genetic diversity within the ssDNA virus group from a broad range of environments and hosts. Novel ssDNA viruses have been discovered in a wide range of plants, animals, fungi, bacteria, Archea and environmental samples. To date members of two families of ssDNA viruses are known to infect plants, the *Nanoviridae* family and the *Geminiviridae* family. Whereas members of four families infect animals, *Anelloviridae*, *Circoviridae* *Parvoviridae* and *Bidnaviridae*, and those of two families infect prokaryotes, *Inoviridae* and *Microviridae*. Only one family is comprised of members which have been identified in Archaea, known as *Spiraviridae*. All these ssDNA viruses have circular ssDNA genomes with the exception of parvoviruses and bidnaviruses which have linear genomes.

In recent years the discovery of several novel viruses which share significant similarities to circoviruses (family *Circoviridae*) has led to the proposal of a new genus known as the cyclovirus genus, which is also falls in the *Circoviridae* family. Cycloviruses were first

discovered in animal faecal samples (Ge *et al.*, 2011; Li *et al.*, 2010; Victoria *et al.*, 2009) and have since been discovered in dragonflies (Dayaram *et al.*, 2013; Rosario *et al.*, 2012a; Rosario *et al.*, 2011), human cerebral fluid and nasopharyngeal aspirates (de Jong *et al.*, 2014; Phan *et al.*, 2014; Smits *et al.*, 2012; van Doorn *et al.*, 2013).

Another group which has been proposed as a new genus is the Gemycircularviruses. Their replication-associated proteins (Reps) share similarities with those encoded by members of geminiviruses and viral sequences integrated in fungal genomes. The first member was isolated from the fungus *Sclerotinia sclerotiorum*, a host in which the virus was shown to induce hypovirulence. Other members of this group have been discovered and are associated with insects, animal faecal material, river sediment and plant material (Dayaram *et al.*, 2012; Du *et al.*, 2014; Kraberger *et al.*, 2013b; Ng *et al.*, 2011b; Sikorski *et al.*, 2013; van den Brand *et al.*, 2012).

Among the recently discovered viruses is a range of novel ssDNA viruses which are yet to be officially indexed at a taxonomic level. A large proportion of these have been discovered using viral metagenomic approaches from environmental samples. Those which do not fall within any of the designated ssDNA families but contain Reps with conserved ssDNA motifs are referred to as circular Rep-encoding single-stranded (CRESS) DNA viruses (Rosario *et al.*, 2012a; Rosario *et al.*, 2012b).

Of all the ssDNA viruses known, the *Geminiviridae* family has been the most studied with more than 300 recognised species and new species being discovered continuously. The large effort towards geminivirus research has primarily been due to the major crop losses they cause and rapid spread as result of their insect vector dynamics.

1.2 Geminiviruses

Geminiviridae is a family of plant infecting viruses. Geminiviruses are found in most regions of the world, with the highest incidence and diversity found within tropical and subtropical regions. These pathogens infect both monocotyledonous (monocot) and dicotyledonous

(dicot) plants and are recognised as a major threat to several economically important crops including tomato, maize, cotton, chickpea and cassava (Varma & Malathi, 2003). Typical disease symptoms can include foliar crinkling, curling, yellowing, stunting, mosaic and/or striations, often resulting in major yield losses.

The circular ssDNA genomes of geminiviruses (~2.7 kilobases (kb) -5.4 kb) are encapsidated in ‘twinned icosahedral’ or ‘geminata’ virions that are ~18-30 nm in size, made up of 22 pentameric capsomers (Fig 1.1) (Zhang *et al.*, 2001). Currently there are seven recognised geminivirus genera: *Begomovirus*, *Curtovirus*, *Topocuvirus*, *Mastrevirus*, *Turncurtovirus*, *Eragrovirus*, and *Becurtovirus* (King *et al.*, 2012; Muhire *et al.*, 2013; Varsani *et al.*, 2014a; Varsani *et al.*, 2014b). Additionally, there are four highly divergent geminivirus which are yet to be classified are Euphorbia caput-medusae latent virus (EcmLV; (Bernardo *et al.*, 2013), Citrus chlorotic dwarf associated virus (CCDaV; (Loconsole *et al.*, 2012) and Grapevine cabernet franc-associated virus (GCFaV; (Krenz *et al.*, 2012) and French bean severe leaf curl virus (unpublished) (Fig. 1.2, 1.3 and Table 1.1). Genera within the *Geminiviridae* family are classified based on insect vector that transmits them, genome organisation and host range (Fauquet & Stanley, 2003)

Begomoviruses can have genomes which consist of either one circular ssDNA component (monopartite) or two circular ssDNA components (bipartite), 2.6-3.2 kb in length, whereas all other geminivirus have been found to be monopartite. Monopartite begomoviruses are often associated with a satellite DNA (alphasatellites and betasatellites), which can increase virulence (Amin *et al.*, 2011; Nawaz-ul-Rehman *et al.*, 2010). A recent study has also shown the mastrevirus species *Wheat dwarf India virus* (WDIV) can also be associated with alphasatellites and betasatellites (Kumar *et al.*, 2014).

Only two genes, those encoding a coat protein (Cp) and the Rep are common to all geminiviruses. Geminivirus genomes encode between four to eight genes depending on species and all have genomes with bidirectionally oriented genes.

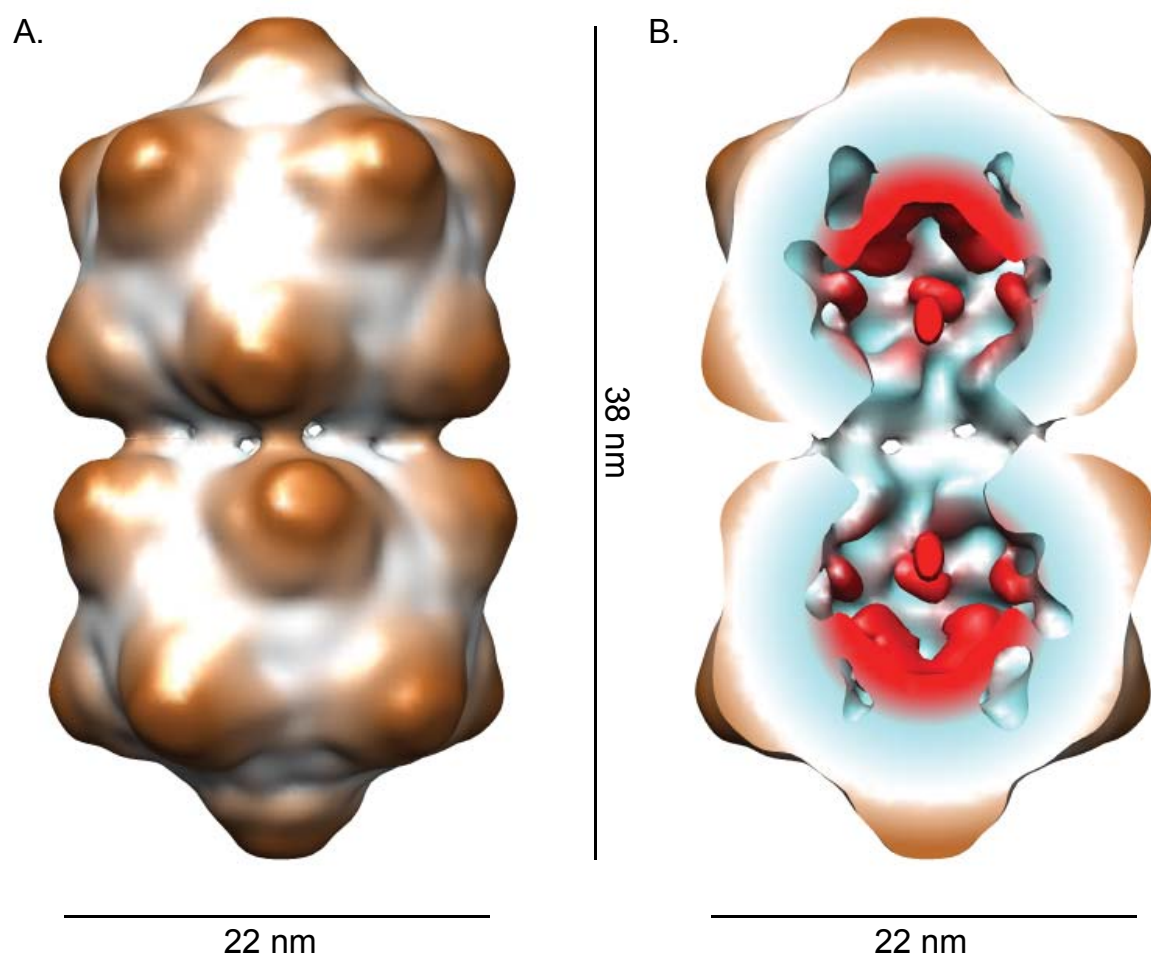


Figure 1.1: Cryo-EM reconstruction of a MSV geminate capsid. A) Outside view of capsid. B) Cross section view of capsid (Shepherd *et al.*, 2010).

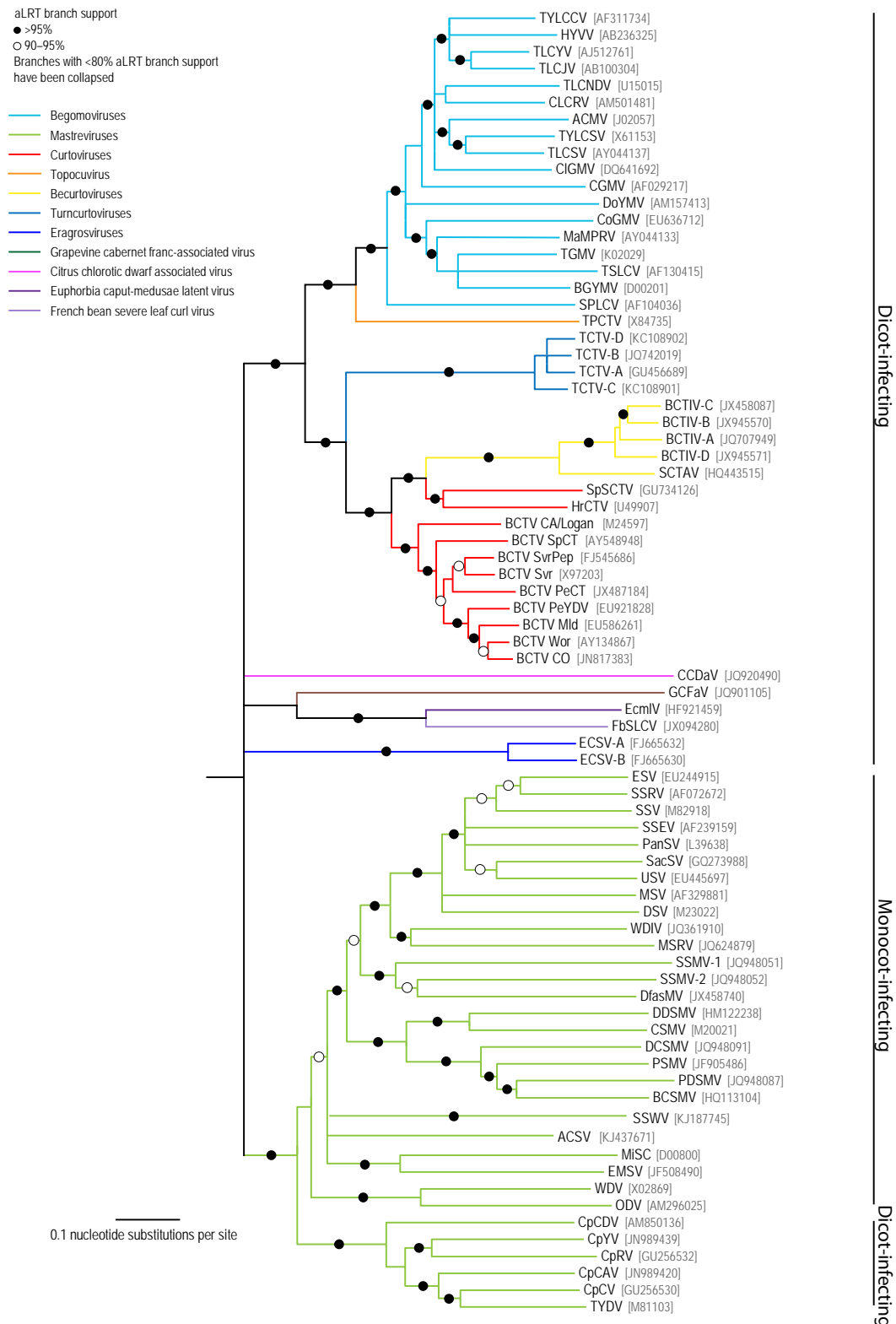


Figure 1.2: Full genome neighbour-joining phylogenetic tree of representative species and strains from geminivirus genera. Viral genomes of geminiviruses across genera are too diverse to be able to credibly align and hence this tree is simply an overview of the phylogenetic relationship between the different genera. Colours shown in key denote each genus and the four unclassified highly diverse geminivirus. aLRT branch support <90% was collapsed. Adapted from Varsani *et al.* (2014b).

Table 1.1 Summary of the key features of the seven classified geminivirus genera and four unclassified highly divergent species.

Genus	No of ICTV approved species	Type species	Natural host range	Vector	Nonanucleotide motif
<i>Begomovirus</i>	~300	<i>Bean golden mosaic virus</i>	Dicotyledon plant <i>sp.</i>	Whitefly	TAATATTAC
<i>Curtovirus</i>	3	<i>Beet curly top virus</i>	Dicotyledon plant <i>sp.</i>	Leafhopper (<i>Circulifer tenellus</i> Baker)	TAATATTAC
<i>Topocuvirus</i>	1	<i>Tomato pseudo-curly top virus</i>	Tomato (<i>Solanum lycopersicum</i>)	Treehopper (<i>Microtalis malleifera</i>)	TAATATTAC
<i>Becurtovirus</i>	2	<i>Beet curly top Iran virus</i>	Dicotyledon plant <i>sp.</i>	Leafhopper	TAAGATTCC
<i>Eragrovirus</i>	1	<i>Eragrostis curvula streak virus</i>	African lovegrass (<i>Eragrostis curvula</i>)	Unknown	TAAGATTCC
<i>Turncurtovirus</i>	1	<i>Turnip curly top virus</i>	Turnip (<i>Brassica rapa</i>), Radish (<i>Raphanus sativus</i>) Diuweed (<i>Descurainia Sophia</i>), (<i>Anchusa sp.</i>), American nightshade (<i>Solanum americanum</i>) and Bladder hibiscus (<i>Hibiscus trionum</i>)	Leafhopper (<i>Cicrulifer haematoceps</i>)	TAATATTAC
<i>Mastrevirus</i>	31	<i>Maize streak virus</i>	<i>Poaceae sp.</i> , <i>Fabaceae sp.</i> and <i>Solanaceae sp.</i>	Leafhopper	TAAT(A/G)TTAC
Unclassified highly divergent Geminivirus					
Citrus chlorotic dwarf associated virus	1	N/A	Citrus (<i>Rutsceae sp.</i>)	Unknown	TAATATTAC
Euphorbia caput-medusae latent virus	1	N/A	Medusa's head (<i>Euphorbia caput-medusae</i>)	Unknown	TAATATTAC
Grapevine cabernet franc-associated virus	1	N/A	Grape (<i>Vitis vinifera</i>)	Unknown	TAATATTAC
French bean severe leaf curl virus	1	N/A	French bean (<i>Phaseolus vulgaris</i>)	Unknown	TAATATTAC

The *Geminiviridae* family is comprised of seven genera and four unclassified divergent viruses, that will most likely be assigned to new genera. A brief descriptive overview of each genus and unclassified divergent species follows (see Fig. 1.2, 1.3 and Table 1.1 for phylogenetic overview, genome organisation, and general information genus within the *Geminiviridae* family, respectively).

1.2.1 The genus *Begomovirus*

The type species of the genus *Begomovirus* is *Bean golden mosaic virus* (and this is one of more than 200 ICTV recognised species (based on a nucleotide sequence identity of <89% for new demarcation of a new species) (Fauquet *et al.*, 2008). This is the only genus within the *Geminiviridae* family that have members that have either bipartite (two ~2.6 kb components) or monopartite (one ~2.7 kb component) genomes. Begomoviruses infect a wide range of dicot plants, however, individual species often have a narrow natural host range (Fauquet *et al.*, 2008). Geographical distribution of begomoviruses extends into the Old World (Africa and Eurasia) and New World (The Americas). Monopartite begomoviruses are thought to have originated in the Old World and are often associated with satellite DNA molecules known as alphasatellites or betasatellites (approximately 1350 nucleotides (nt) in size) and can affect pathogenicity and symptomology in the host (Zhou, 2013). Alphasatellites encode a single protein which is a Rep that is most closely related to the Rep encoded by members of the *Nanoviridae* family. These are capable of self-replicating but need an associated begomovirus for encapsidation and movement. Alphasatellites and betasatellites have been shown to interfere with RNA silencing by targeting a step in the RNA-silencing pathway (Amin *et al.*, 2011; Nawaz-ul-Rehman *et al.*, 2010). Betasatellites share a nonanucleotide motif sequence (TAATATTAC) with helper begomovirus and encodes a single gene which is a pathogenicity determinant. Although satellites are not usually associated with or needed for increased virulence in bipartite begomoviruses, some have been found co-infecting the same host (Mansoor *et al.*, 2003). All begomoviruses have a nonanucleotide motif sequence of “TAATATTAC”. Monopartite begomovirus genomes encode six genes and bipartite begomovirus encode six to eight genes. The genome organisation of monopartite and DNA-A of bipartite begomoviruses is similar, both encoding the following genes: coat protein gene (*cp*, V1), replication-associated protein gene (*rep*, C1), transcriptional activator protein gene (*trap/ss*, C2), a replication enhancer gene (*ren*, C3) and a symptom determinant gene (*sd*, C4) (Fig. 1.3). In addition, monopartite and DNA-A

components of Old World begomoviruses also encode a pre-coat gene (V2). The DNA-B of bipartite begomoviruses encodes a nuclear-shuttle protein (*nsp*, V1) and a movement protein (*mp*, C1). All begomoviruses are transmitted by the whitefly *Bemisia tabaci*, with the exception of most *Abutilon mosaic virus* strains which have been vegetatively propagated in ornamental plants for a long period of time and a mutation in the CP has rendered this virus unable to be transmitted by whiteflies (Höhnle *et al.*, 2001).

1.2.2 The genus *Curtovirus*

The *Curtovirus* genus is comprised of three species, the type species *Beet curly top virus* (BCTV), *Spinach severe curly top virus* (SpSCTV) and *Horseradish curly top virus* (HrCTV) a taxonomy classification that was recently revised by Varsani *et al.* (2014a), as determined by a nucleotide sequence identity of <77% for demarcation of a new species. Members of this genus are known to infect more than 300 dicot plant species and classical symptoms include leaf curling and distortion, yellowing of leaves, vein swelling, stunting and necrosis. To date curtoviruses have only been identified in the northern hemisphere (Baliji *et al.*, 2004; Creamer *et al.*, 2005; Velásquez-Valle *et al.*, 2012). Many isolate sequences of BCTV have been deposited in GenBank, with a total of 9 strains identified (Briddon *et al.*, 1998; Chen *et al.*, 2011; Hormuzdi & Bisaro, 1993; Lam *et al.*, 2009; Stanley *et al.*, 1986; Stenger, 1993; Varsani *et al.*, 2014a). Curtoviruses have a nonanucleotide motif sequence of “TAATATTAC” and can encode up to seven genes (BCTV, Fig. 1.3). Genes encoded include three on the virion-sense strand, *mp* (V2), *reg* (V3) and the *cp* (V1), and four on the complementary-sense strand, the *rep* (C1), *ren* (C2), *trap/ss* (C3) and *sd* (C4). The complementary-sense genes are most similar to that of the begomoviruses and those in the virion-sense to other monopartite geminiviruses. Curtoviruses are transmitted by the leaf hopper *Circulifer tenellus* Baker (Chen & Gilbertson, 2008).

1.2.3 The genus *Topocuvirus*

The sole member of the genus *Topocuvirus* is the species *Tomato pseudo-curly top virus* (TPCTV). A single isolate was recovered from a tomato (*Solanum lycopersicum*) plant collected in Florida presenting leaf curling symptoms (Briddon *et al.*, 1996). The symptoms presented are very similar to those caused by BCTV, which lead to the initial thought that this was the cause of the disease. Subsequent discovery that BCTV was not the causal agent the

disease led to this virus being named pseudo-curly top which was soon followed by characterisation of the TPCTV. TPCTV has a nonanucleotide motif sequence of “TAATATTAC”. Genome organisation consists of six open reading frames (ORFs), two virion-sense ORFs encoding a *mp* (V2) and a *cp* (V1), and four complementary-sense, a unknown (C4), *rep* (C1), possible *trap/ss* (C2) and a *ren* (C3) (Fig. 1.3). The vector of this virus is most likely the treehopper species *Micrutalis malleifera* which was shown to be able to transmit the disease agent causing pseudo-curly top disease in tomato before the virus was molecularly characterised (Simons & Coe, 1958).

1.2.4 The genus *Becurtovirus*

The *Becurtovirus* genus name is derived from the first species identified *Beet curly top Iran virus* (BCTIV) (Heydarnejad *et al.*, 2013; Soleimani *et al.*, 2013; Varsani *et al.*, 2014b; Yazdi *et al.*, 2008). The other recognised species is *Spinach curly top Arizona virus* (SCATV) (Hernández-Zepeda *et al.*, 2013). So far BCTIV has only been found in Iran infecting dicot plants (Gharouni Kardani *et al.*, 2013; Heydarnejad *et al.*, 2013; Yazdi *et al.*, 2008), and SCATV in the United States of America infecting spinach (*Spinacia oleracea*) (Hernández-Zepeda *et al.*, 2013). The nonanucleotide sequence for all members of this genus is “TAAGATTCC”, which is different from the conserved sequence seen among most geminivirus. Becurtoviruses have five ORFs, three virion sense ORFs and two complement sense ORFs (Fig. 1.3). Those that are virion sense oriented, encode for a *mp*, possible *reg* and a *cp*, and are most closely related to those in a similar position in the curtoviruses. The complementary sense ORFs encode for a *rep* (mostly likely derived from a spliced transcript) and a *repA*. Symptoms in infected hosts include leaf deformation, rolling, yellowing, stunting and vein swelling (Gharouni Kardani *et al.*, 2013; Hernández-Zepeda *et al.*, 2013). BCTIV is transmitted by the leaf hopper species *Circulifer haematocephus* (Heydarnejad *et al.*, 2013). The vector for SCATV is still unknown however it is thought to most likely be the leaf hopper species *C. tenellus* as this is responsible for spreading the Curtovirus species SpSCTV, which was found co-infecting spinach plants along with SCATV.

1.2.5 The genus *Eragrovirus*

The *Eragrovirus* genus thus far consists solely of the species *Eragrostis curvula streak virus* (ECSV) (Varsani *et al.*, 2009b). There are two strains ECSV-A and -B. ECSV has so far only been found in South Africa infecting the perennial grass *Eragrostis curvula* which presented

mild streak symptoms, similar to that caused in maize infected with *Maize streak virus* (MSV). Members of this species all have an atypical nonanucleotide sequence motif of “TAAGATTCC”, the same motif which is seen in the becurtoviruses. Based on current annotation of ECSV genome encodes for four ORFs, two transcripts in each orientation (Fig. 1.3). On the complementary-sense strand are two ORFs encoding the *rep* and possible *trap* and on the virion-sense strand are two ORFs encoding a possible *mp* and a *cp*. The Rep and CP both share similarity to other geminiviruses, whereas the possible MP and Trap share no homology to other geminiviruses, the latter do however correspond in positioning in genome to other geminivirus (Varsani *et al.*, 2014b; Varsani *et al.*, 2009b). Based on phylogeny the Rep is most closely related to begomoviruses, topocuviruses and curtoviruses whereas the CP is most closely related to the mastreviruses, alluding to the fact that this recently proposed genus is the progeny of an ancient recombination event.

1.2.6 The genus *Turncurtovirus*

Turncurtovirus is another genus which only has a single recognised species *Turnip curly top virus* (TCTV) (Briddon *et al.*, 2010a; Razavinejad & Heydarnejad, 2013; Razavinejad *et al.*, 2013). There are currently four strains of TCTV (A – D) (Razavinejad *et al.*, 2013). TCTV has been found only in Iran and through molecular methods (full genome isolation or screening PCR) a broad host range has been identified including *Brassica rapa*, *Raphanus sativus*, *Descurainia sophia*, *Anchusa sp.*, *Solanum americanum* and *Hibiscus trionum* (Briddon *et al.*, 2010a; Razavinejad *et al.*, 2013). Symptoms in turnips include leaf cupping and vein swelling. The nonanucleotide motif sequence is the characteristic “TAATATTAC” present in the majority of geminiviruses. The TCTV genome has six ORFs, similar to some begomoviruses. Two ORFs in the virion-sense encode a possible *mp* (V2) and a *cp* (V1) and four ORFs in the complementary-sense encode a possible *sd* (C4), *rep*, possible *trap/ss* (C2) and a *ren* (C3). The genome organisation TCTV is more similar to topocuviruses, however, its biological properties are more similar to curtoviruses. This virus is vectored by the leaf hopper species *C. haematoceps* (Razavinejad & Heydarnejad, 2013).

1.2.7 The genus *Mastrevirus*

The type member of the *Mastrevirus* genus is *Maize streak virus*. MSV particles were first visualised in 1974 (Bock *et al.*, 1974). Mastreviruses infect either mono or dicot plant species and have been found only in the old world. An exceptions to this are three unclassified

Mastrevirus species, two of which were isolated from dragonflies collected in Puerto Rico, with the nonmonoclature Dragonfly-associated mastrevirus (Rosario *et al.*, 2013) and a mastrevirus-like sequences identified in sweet potato samples from Peru (Kreuze *et al.*, 2009). Thirty-two species constitute the *Mastrevirus* genus, five of which were discovered in studies undertaken as part of this thesis, see Chapter 2 and 5 (based on a nucleotide sequence identity of <78% for new demarcation of a new species) (Ali *et al.*, 2004; Briddon *et al.*, 2010b; Chatani *et al.*, 1991; Geering *et al.*, 2011; Greber, 1989; Hadfield *et al.*, 2011; Hadfield *et al.*, 2012; Halley-Stott *et al.*, 2007; Krabberger *et al.*, 2012; Kumar *et al.*, 2012; Lawry *et al.*, 2009; MacDowell *et al.*, 1985; Martin *et al.*, 2001; Nahid *et al.*, 2008; Oluwafemi *et al.*, 2008; Pande *et al.*, 2012; Rybicki, 1994; Schnippenkoetter *et al.*, 2001; Schubert *et al.*, 2007; Shepherd *et al.*, 2008b; Thomas *et al.*, 2010). Until recently the only geminivirus genus known to be associated with an alphasatellite or a betasatellite was begomoviruses, Kumar *et al.* (2014) have identified both types of satellites associated with the mastrevirus species WDIV. Mastreviruses have the nonanucleotide motif sequence “TAAT(A/G)TTAC”. All species have four ORFs, with a *mp* (V2) and a *cp* (V1) encoded on the virion-sense strand and a *rep* and *repA* encoded on the complementary-sense strand (Fig. 1.3). The *rep* is expressed from a spliced ORF C1 and C2 and *repA* is expressed from ORF C1 alone (Dekker *et al.*, 1991; Mullineaux *et al.*, 1990; Schalk *et al.*, 1989; Wright *et al.*, 1997). All four ORFs are necessary for systemic infection (Liu *et al.*, 1998). The different mastrevirus members are each transmitted by leafhopper species from the family *Cicadellidae*.

1.2.8 Unclassified highly divergent geminivirus

Four highly divergent, yet to be classified, geminiviruses are EcmLV (HF921459, HF921460 and HF921477; (Bernardo *et al.*, 2013), CCDaV (JQ920490 and KF561253; (Loconsole *et al.*, 2012), GRLaV (KC427993-96; (Krenz *et al.*, 2012; Poojari *et al.*, 2013) and FbSLSV (JX094280-81; unpublished) (Fig. 1.2). EcmLV was isolated from *Euphobia caput-medusa*, a host which was non-symptomatic, CCDaV from several citrus *sp.*, GRLaV from grapevine and FbSLSV from a common french bean. All have the highly conserved geminivirus nonanucleotide motif “TAATATTAC” and all potentially have a spliced *rep*. Full genome comparison shows FbSLSV and EcmLV are most closely related to each other (~72% nucleotide pairwise identity). EcmLV and FbSLSV have similar genome organisation however EcmLV has seven ORF's and FbSLSV only has six. Both have three ORFs in the

complementary-sense orientation, which encode a spliced *rep* (C1:C2), a *repA* (C1) and an unknown ORF. In the virion-sense orientation FbSLSV has three ORFs encoding; a *cp* (V1) and two unknown (V2 and V3), whereas EcmLV has four ORFs encoding a *cp* (V1), a possible *mp* (V4 and/or V3) and an unknown expressed from a spliced transcript (V3 and V4). CCDaV which contains five ORFs, in the complementary-sense orientation has a spliced *rep* (C1:C2) and a *repA* (C1), and in the virion-sense a *cp* (V1) and possible *mp* (V4 and/or V3). Lastly GCFaV has six ORFs, in the complementary-sense orientation, a spliced *rep* (C1:C2), a *repA* (C1) and an unknown (C3), in the virion-sense orientation, a *cp* (V1) and two unknown (V2 and V3) ORFs. The Reps of FbSLSV and EcmLV are most closely related to GCFaV (~79% amino acid pairwise identity) (Bernardo *et al.*, 2013). Full genome nucleotide analysis of GCFaV indicates it is most closely related to the mastreviruses and ECSV (~50% nucleotide pairwise identity). CCDaV has a genome of ~3640 nt, which is the largest genome of all geminiviruses.

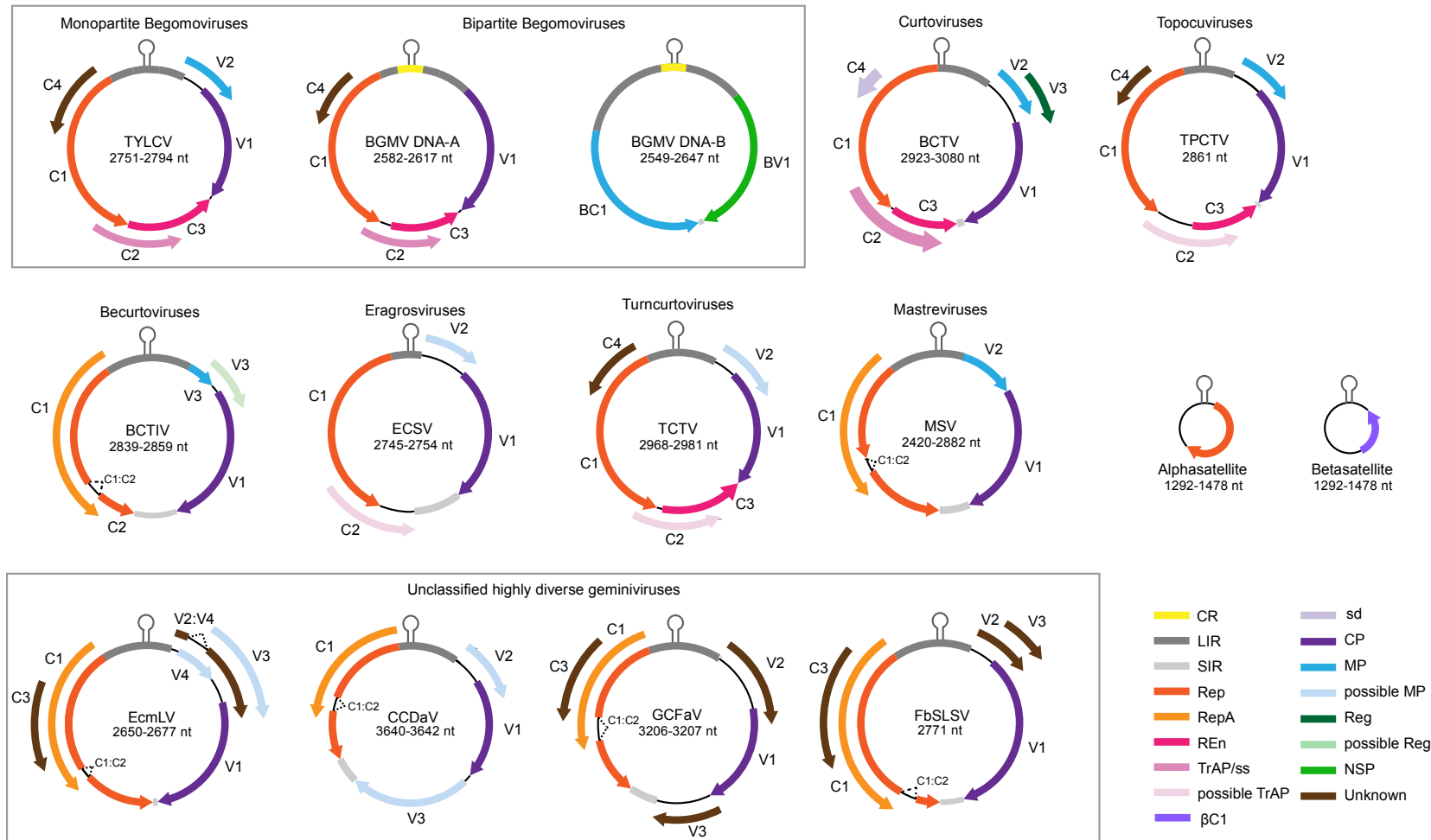


Figure 1.3: Genome organisations of representatives from each geminivirus genus and four highly divergent uncharacterised geminiviruses. Arrows denote direction in which each ORF is transcribed. The protein encoded by each ORF is represented by colours shown in key. The following are acronyms and genes or elements they represent, CR (common region), LIR (long intergenic region), SIR (short intergenic region), *rep* (replication-associated protein), *ren* (replication enhancer gene), *trap* (transcription activator protein gene), *ss* (silencing suppressor), *sd* (symptom determinant), *cp* (coat protein), *mp* (movement protein), *reg* (regulatory gene) and *nsp* (nuclear shuttle protein).

1.2.9 Replication of geminiviruses

Geminiviruses replicate through a mechanism known as rolling circle replication (RCR) (Fig 1.4) (Heyraud *et al.*, 1993; Jeske *et al.*, 2001; Koonin & Ilyina, 1992; Saunders *et al.*, 1991) and the less studied recombination-dependent replication (RDR) (Jeske *et al.*, 2001; Saunders *et al.*, 1991). Geminivirus replication occurs in the plant host cell nucleus through a double stranded DNA (dsDNA) intermediate. The small genome size of these viruses means their genomes encode for only a few proteins and do not encode their own DNA polymerases, therefore relying largely on the plant host replication machinery and cell cycle for its own replication. RCR in geminiviruses can be broken down into stages referred to as initiation, elongation and termination; reviewed by (Gutierrez, 1999; Hanley-Bowdoin *et al.*, 2013; Martin *et al.*, 2011a). The following is a collective overview of RCR of geminiviruses through experimental information from a variety of geminivirus species; however it is worth noting that this may not be the exact series of events for all geminiviruses. Following entry into the cell via the insect vectors stylet the viral DNA is transported to the infected cell nucleus presumably by the host transport systems. Several lines of evidence show that the CP is involved in the localisation of viral DNA to host nucleus (Liu *et al.*, 2001; Liu *et al.*, 1999a; Qin *et al.*, 1998; Unseld *et al.*, 2001), in mastreviruses it has been shown that the CP also enters the nucleus (Liu *et al.*, 2001). Details of when and how the viral capsid is shed prior to replication are still to be fully elucidated. Once in the nucleus, the viral ssDNA becomes covalently closed circular double-stranded DNA (cccdsDNA) by the host DNA polymerase. This is primed either by a sequence of oligonucleotides of host origin or in the case of mastreviruses a primer exists which is annealed to the parent viral DNA and encapsidated (Andersen *et al.*, 1988; Donson *et al.*, 1984; Hayes *et al.*, 1988). Following conversion into cccdsDNA, this molecule most likely associates with host histone proteins and is packaged into mini-chromosomes ready for gene transcription. Expression of the Rep protein is crucial for initiation of RCR (Gröning *et al.*, 1990; Pilartz & Jeske, 1992; Saunders *et al.*, 1991). To then initiate RCR replication the Rep protein binds to repeating *cis*-acting elements known as “iterons” (Argüello-Astorga & Ruiz-Medrano, 2001; Londoño *et al.*, 2010), in close proximity to the stem-loop which contains the highly-conserved nonanucleotide 5'-TAATATTAC-3,' known as the origin of replication (*v-ori*) (among geminivirus species there are a few variations of the nonanucleotide) and Rep cleaves the positive viral DNA strand. Viral DNA in an open circular state becomes a template for a continuous coil-like production of a new virion DNA strand while displacing the old strand.

Following one or more cycles the ends of the old virion DNA strand are completely displaced and subsequently ligated, resulting circular virion ssDNA can consist of a single copy or several copies of the original circular ssDNA molecule. If monomeric, these molecules can either then be encapsidated into virions or re-enter the replication cycle.

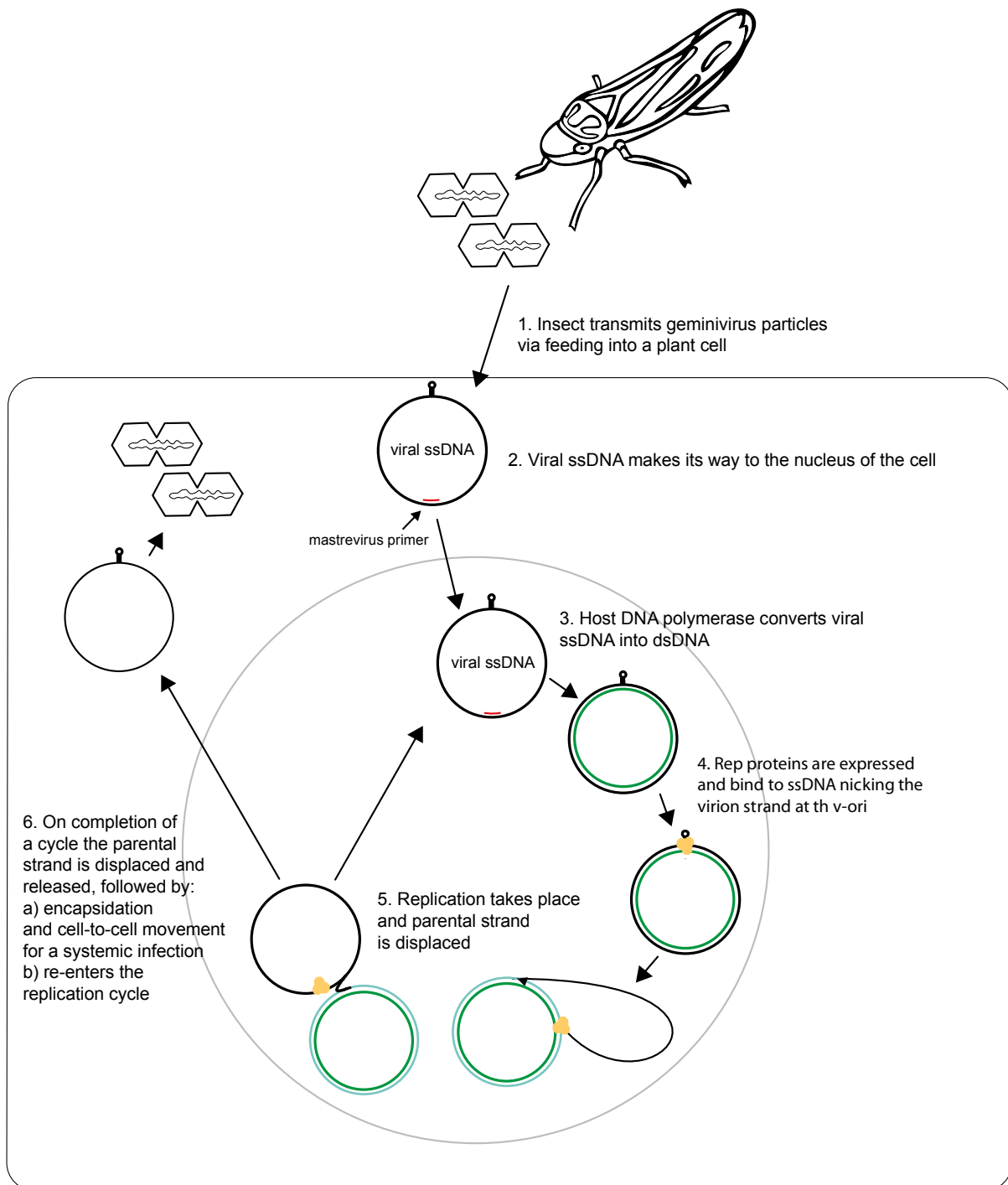


Figure 1.4: Summary of rolling circle replication of mastreviruses. Following entry into the cell via a feeding leafhopper (step 1), viral DNA is transported to the nucleus (step 2) where it is converted to dsDNA, a replicative intermediate form (step 3). The Rep is transcribed and binds to the *v-ori* where it nicks the virion strand (step 4). The Rep remains bound to the nicked end while the new strand is synthesised by the host replication machinery (step 5). Once a cycle of replication is completed the parental strand is displaced (step 6) and released this can then be encapsidated and move cell-to-cell to establish a systemic infection (step 6A) or re-enter the replication cycle in increase viral load in the cell (step 6B).

1.2.10 Theories behind the evolutionary origin of geminiviruses

A common hypothesis regarding the origin of viruses is that they were present in the early stages of life on earth and possibly even pre-dated the divergence of cellular life (Forterre, 1992; Forterre, 2006; Koonin *et al.*, 2006). One of the major roots of this theory is that viruses which infect a range of hosts from the three domains of life often share genomic and morphological features. Other discussed theories are; 1) viruses were formed from unicellular organisms as part of a reduction evolution and 2) they originated from genetic material which separated from its cellular host during replication or other cell cycle stages and then became parasitic (Hendrix *et al.*, 2000). It is also possible that the emergence of viral families was polyphyletic.

The origin of geminiviruses is a well theorised topic. Like geminiviruses, several families of circular ssDNA viruses replicate through RCR. These includes families which infect plants (*Nanoviridae*), animals (*Circoviridae*) and bacteria (*Microviridae*) because these viruses most likely replicate in a similar manner and share homologous genetic features, mostly in the Rep, which suggests they may have originated from a common ancestor (Ilyina & Koonin, 1992; Koonin *et al.*, 2006; Rojas *et al.*, 2005). Two lines of evidence lead to the theory that geminiviruses originated from bacterial plasmids, referred to as the plasmid-to-virus hypothesis (Krupovic *et al.*, 2009); 1) Geminiviruses are able to replicate efficiently in *Agrobacterium tumefaciens* and to some degree in *Escherichia coli* (Rigden *et al.*, 1996; Selth *et al.*, 2002). Showing that bacterial cell cycle factors can be hijacked by geminiviruses for viral replication. 2) Identification of ssDNA Rep-like sequences in gram-positive bacteria and bacterial plasmids known as extrachromosomal DNA replicons (EcDNA) (Krupovic *et al.*, 2009). Specifically plasmids from the insect transmitted plant pathogens phytoplasmas (Nishigawa *et al.*, 2001; Nishigawa *et al.*, 2002; Oshima *et al.*, 2001; Rekab *et al.*, 1999; Tran-Nguyen & Gibb, 2006). Krupovic *et al.* (2009) also discuss the observation that the CP of geminiviruses has similar features to ssRNA viruses, in particular *Satellite tobacco necrosis virus*. Therefore the proposed simplified scenario is that ssRNA viruses and phytoplasma occupied the same host and a recombination event occurred between the phytoplasma plasmid and ssRNA virus for acquisition of a CP, ultimately leading to the emergence of early geminiviruses.

This hypothesis was rejected by Saccardo *et al.* (2011), based on evidence that the Rep features in EcDNA are not of phytoplasmal origin and ancestral geminivirus-like CPs have been found in viruses from marine environments in which similarity analysis shows is a more likely origin than a recombination event with a ssRNA virus. There are actually three types of EcDNA associated with phytoplasma, two of which type I and II replicate through RCR. Type I is more closely related to plasmids which employ RCR from the pLS1 family (Bergemann *et al.*, 1989). Type II is the only group which shares similarity to the geminivirus Repls, evidence given by Saccardo *et al.* (2011) indicates that the original type II Rep was swapped through recombination for a Rep from a geminivirus-like replicon. Additionally geminivirus Rep-like protein sequences have been identified in eukaryotic genomes; plants (Kenton *et al.*, 1995; Murad *et al.*, 2004), fungi and entomoeba. Phylogenetically, the Rep-like sequences of plant origin cluster more closely to those of fungal origin (Liu *et al.*, 2011). Integrated geminiviral DNA has been identified in plant genomes, for example ancient germplasm of *Nicotiana sp.* with integrated geminivirus-related DNA has led to the suggestion that geminiviruses probably originated >10 million years ago (Ashby *et al.*, 1997; Bejarano *et al.*, 1996; Lefeuvre *et al.*, 2011; Lim *et al.*, 2000; Murad *et al.*, 2004).

A ssDNA virus which infect fungi, *Sclerotinia sclerotiorum* hypervirulence-associated DNA virus 1, has a Rep which is most closely related to those of geminiviruses. This virus is member of the proposed Gemycircularviruses genus (Rosario *et al.*, 2012a; Sikorski *et al.*, 2013; Yu *et al.*, 2010; Yu *et al.*, 2013). It is believed that these two groups, the geminiviruses and gemycircularviruses most likely evolved independently but may have a common ancestor which existed before the split off between plants and fungi (Liu *et al.*, 2011).

1.2.11 Geminivirus evolution

1.2.11.1 Genetic drift

For some time it was thought that RNA viruses evolve much faster than DNA viruses, However, several studies have shown that small ssDNA viruses are able to evolve at rates similar to those of RNA viruses, and used this information to date and ascertain possible origins of geminiviruses (De Bruyn *et al.*, 2012; Duffy & Holmes, 2008; 2009; Duffy *et al.*,

2008; Firth *et al.*, 2009; Grigoras *et al.*, 2010; Harkins *et al.*, 2014; Harkins *et al.*, 2009b; Krabberger *et al.*, 2013a; Van der Walt *et al.*, 2008a).

Mutation rates in geminiviruses has been shown to be relatively high and comparable to those determined for RNA viruses (Duffy & Holmes, 2008; 2009; Duffy *et al.*, 2008; Roossinck, 1997). For example, mastreviruses have been shown to have substitution rates of between 2 and 3×10^{-4} substitutions/site/year (Harkins *et al.*, 2009a). It is surprising that this is the case as geminiviruses are replicated by a presumed high fidelity host polymerase which has a much lower error rate than the error prone RNA-polymerase encoded by RNA viruses. There are several possible explanations for this, such as the host exonucleases may not repair errors in the viral DNA because it is not methylated unlike the plant host DNA or it may be that the virus is double stranded for a short period during replication. Mutation rates can also be affected by base deamination which is the spontaneous conversion of a cytosine (C) to a uracil and a guanine (G) to a xanthine (Caulfield *et al.*, 1998; Duffy & Holmes, 2008; Inamdar *et al.*, 1992). This is more likely to occur when DNA is in a single-stranded form for long periods of time prior to replication (Duffy & Holmes, 2008). It is unknown to what degree deamination affects mutation rates.

Van der Walt *et al.* (2008a) investigated mutation rates on both strands to identify substitution biases in MSV and found that G to thymine (T) mutations were most common across the whole genome. It was proposed that the cause of this is not due to deamination but possibly oxidation of guanine. Oxidation is a defence mechanism employed by the host (Van der Walt *et al.*, 2008a). Furthermore substitution biases seem to be strand specific with C to adenine (A) mutations most commonly occurring on the complementary strand whereas G to T mutations most commonly occurring on the virion strand. Recombination has been affiliated with mutational changes in ssDNA genomes (Shcherbakov *et al.*, 2011) however, in geminiviruses, controlled short term experiments showed no association (Monjane *et al.*, 2012).

Both *Tomato yellow leaf curl virus* (TYLCV) and MSV seem to be evolving predominantly under negative selection (Duffy & Holmes, 2008; Monjane *et al.*, 2012; Van der Walt *et al.*, 2008a). Data on geminivirus mutation rates from both field and lab experiments as well as

those from different natural hosts are suggesting that selection pressures may not, to a large degree, be influencing substitution rates (Isnard *et al.*, 1998; Monjane *et al.*, 2012; Van der Walt *et al.*, 2008a). Harkins *et al.* (2009a) showed through long-term evolution experiments that mastrevirus are under a similar level of negative pressure to other geminiviruses and therefore rejects the hypothesis put forward by Wu *et al.* (2008) that mastrevirus coevolved with their hosts. In order to investigate geminiviral evolution and timelines it is necessary to estimate basal substitution rates. Sampling of geminivirus isolates in nature at various time points such as those from archived samples with known collection dates has enabled the fairly accurate estimation of some geminivirus timelines. Additionally, this information has enabled studies to infer the possible geographical origin of various geminivirus species most recent common ancestor (MRCA), with most of these studies focusing on begomovirus (De Bruyn *et al.*, 2012; Lefeuvre *et al.*, 2010) and mastrevirus species (Harkins *et al.*, 2009b; Krabberger *et al.*, 2013a; Monjane *et al.*, 2011b).

1.2.11.2 Recombination and reassortment

Recombination is the swapping of genetic material from one virus to another, a mechanism which accelerates geminivirus evolution. Recombination of eukaryote-infecting ssDNA viruses has been reviewed by Martin *et al.* (2011a). It has also been extensively analysed in geminiviruses and evidence of recombination has been identified amongst strains, species and even between genera of geminiviruses. Bipartite geminiviruses pseudo-recombination (also known as reassortment) is the swapping of entire components among strains or species, has also been documented. An obvious prerequisite for recombination to occur is that two viruses co-infect the same cell. Recombination enables the repair of deleterious mutations and can preserve positive mutations in a population. There is homologous recombination which can allow two closely related defective viral isolates to form a “fitter” virus and non-homologous recombination that can enable the shuffling of viral genes as well as insertion of new genes or intergenic regions. Recombination detected in most geminiviruses shows that there are two recombination hotspots, one at the interface between the CP and SIR, and the other in the LIR close to the *v-ori* (Lefeuvre *et al.*, 2009; Martin *et al.*, 2011b; van der Walt *et al.*, 2009; Varsani *et al.*, 2009a; Varsani *et al.*, 2008b).

The mechanisms facilitating recombination are still not very well understood, however, a common theory is that it may occur following the disruption of replication when transcription

and replication factors collide. This is supported by recombination hotspots at the interfaces between ORF and intergenic regions as discussed earlier (Jeske *et al.*, 2001). Premature detachment of the replication complex or various other factors may also result in the reattachment of to a new viral template to produce a recombinant. Secondary structures are most likely a contributing attribute to facilitating recombination as they will likely cause the replication machinery to stall, reinforced by the evidence of a recombination hotspot close to the *v-ori*, which forms a hairpin. It is also possible that while the virus is in a dsDNA replicative state, a DNA breakage may occur or it may be in the covalently closed or open circular form and the host repair systems could match this template strand with strand from another similar virus. The likelihood of two strands being repaired will most likely depend on the level of similarity shared in the repair region. This mechanism is known as “recombination dependent replication” which results in high frequencies of sub-genomes length DNA molecules and been shown to be a mode of viral genome replication in geminiviruses (Alberter *et al.*, 2005; Casado *et al.*, 2004; Erdmann *et al.*, 2010; Jeske *et al.*, 2001; Martin *et al.*, 2011a; Preiss & Jeske, 2003).

By recombining, geminiviruses are able to rapidly explore sequence space and therefore may have the ability to adapt to new hosts and vectors more quickly than through mutation alone. The ability to adapt rapidly through recombination has been demonstrated in controlled lab experiments using chimeric defective MSV clones constructed from wild-type “fit” MSV genome. Combining these defective MSV genomes in a mixed infection resulted in a rich tapestry of recombinant progeny (van der Walt *et al.*, 2009) and in an experiment where the genomes were originally detrimentally defective, these viruses were able to recombine to produce viable progeny (Monjane *et al.*, 2014). Studies of geminiviruses outside of the controlled lab experiments have implicated recombination as a possible driving force behind increase pathogenicity, host range and emergence of new geminiviruses (Klute *et al.*, 1996; Monjane *et al.*, 2011b; Padidam *et al.*, 1999; Pita *et al.*, 2001; Ribeiro *et al.*, 2003; Sanz *et al.*, 2000; Saunders *et al.*, 2002; van der Walt *et al.*, 2009; Varsani *et al.*, 2008b; Zhou *et al.*, 1997; Zhou *et al.*, 1998).

1.2.12 Genome secondary structure

Geminivirus single-stranded genomic DNA forms secondary structures which play important roles in replication and potentially other biological functions. The most studied secondary

structure in geminivirus and other circular ssDNA viruses is the stem-loop structure which contains the *v-ori* central to replication (Muhire *et al.*, 2014; Orozco & Hanley-Bowdoin, 1996; Stanley, 1995). A recent review looked at secondary structures across geminivirus in order to identify if any other highly conserved secondary structures could potentially play a role in biological functions of the virus (Muhire *et al.*, 2014). This study identified three highly conserved secondary structures. The first of which is located in the Rep intron which has been previously reported in MSV by Shepherd *et al.* (2006). The second identified structure is associated with the *mp* intron of mastrevirus, this structure is suggested to play a role in splicing of the *mp*. The third structure was identified near the end of the *cp* gene in begomovirus and encompasses the predicted polyadenylation signals for transcription of both strands therefore making it a likely element associated with transcription factors. It is likely that there are many undocumented secondary structures throughout the geminivirus genome, which play a part in influencing biological features which we are yet to elucidate.

1.3 Mastreviruses

1.3.1 Genomes of mastreviruses

The mastrevirus genomes consists of a LIR and SIR and four ORFs (Fig. 1.3 and 1.5). The following is a summary of the molecular features in the mastrevirus genomes and functions the proteins encoded. Several reviews discuss geminiviral protein structure and function (Boulton, 2002; Fondong, 2013; Gafni & Epel, 2002; Gutierrez *et al.*, 2004). Two mastreviruses, *Wheat dwarf virus* (WDV) and *Maize streak virus* (MSV) have been extensively studied and hence the following summary is based on them. Although there are differences amongst mastreviruses the key features are most likely very similar.

1.3.2 Intergenic regions

Located within the SIR is the complementary strand origin of replication site, initiating negative viral DNA strand synthesis, this site it is primed by a tightly bound ~80 nt primer which is encapsidated along with the viral DNA (Donson *et al.*, 1984; Hayes *et al.*, 1988; Kammann *et al.*, 1991) (Fig. 1.5). The other main features within SIR are the termination and polyadenylation signals for complementary-sense and large virion-sense transcripts (Fig. 1.5).

The LIR is the more studied intergenic region of the two and has several important features which include the *v-ori*, viral *rep* and host protein recognition/binding sites as well as virion-sense and complementary-sense promoter sites and transcription regulatory elements. Initiation of RCR commences following the nicking of DNA at the *v-ori* which is in the conserved nonanucleotide motif TAAT(A/G)TTAC between nucleotides 7 and 8. This motif is in the loop of a stem-loop which is situated approximately in the middle of the LIR in mastreviruses.

In geminiviruses iterative *cis*-acting sequence elements close to the *v-ori* and upstream of the *rep* gene TATA evidently act as specific binding or recognition sites for the *rep* in order to commence RCR (Argüello-Astorga *et al.*, 1994; Argüello-Astorga & Ruiz-Medrano, 2001; Castellano *et al.*, 1999; Fontes *et al.*, 1992; Orozco *et al.*, 1998; Orozco & Hanley-Bowdoin, 1998). These iterative elements are referred to as iterons. Several studies highlight potential iterons in begomovirus and *rep* interactions with these sites support a similar interaction in mastreviruses (Sanz-Burgos & Gutiérrez, 1998; Willment *et al.*, 2007). Unlike in begomoviruses where these iterons are conserved, they are highly variable in mastreviruses (Argüello-Astorga *et al.*, 1994). Nonetheless, Willment *et al.* (2007) demonstrated that Reps of different mastrevirus species can replicate each other with a relatively high level of efficiency. These authors concluded that although iterons are important Rep recognition/binding sites some variability in sequence is acceptable in order for adequate levels of replication to occur.

First discovered in WDV is a sequence consisting of a series of A-T's, found between the stem-loop and beginning of the *mp* ORF. This A-T sequence series is an intrinsic bent DNA region in mastreviruses proposed to be a replication regulatory element (Gutiérrez *et al.*, 1995; Suárez-López *et al.*, 1995). Upstream of the stem-loop is a GC-rich region (GC boxes) which binds host nuclear factors shown *in vitro* by Fenoll *et al.* (1990). Both intergenic regions play an important role in viral replication and expression of viral genes.

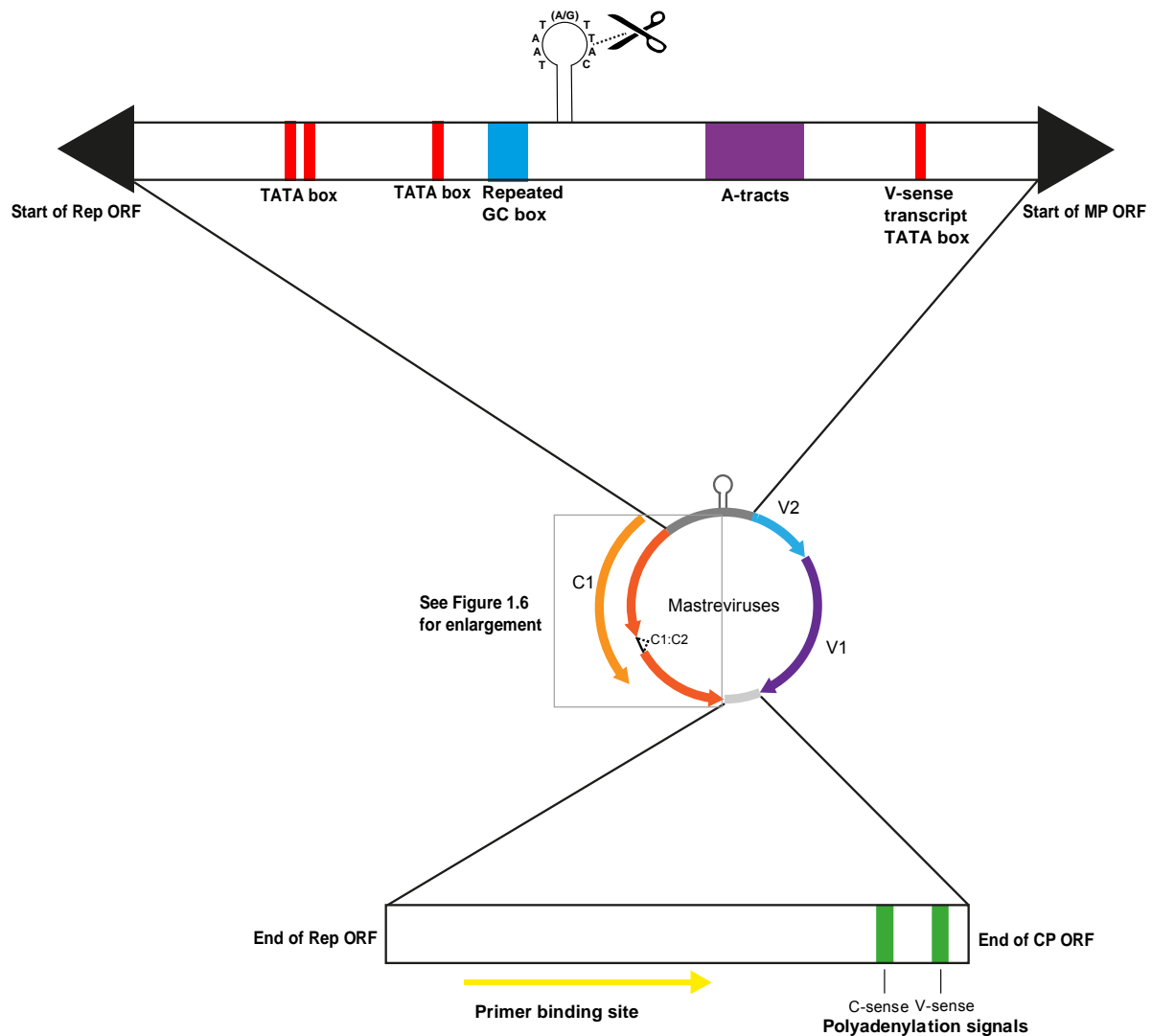


Figure 1.5: Overview of functional regions within the long intergenic (LIR) and short intergenic regions (SIR) of mastrevirus (based on WDV and MSV). The key feature identified in the LIR are the stemloop containing the nicking site for initiation of rolling circle replication, a GC box which is a repetitive sequence shown to bind host factors (Fenoll *et al.*, 1990) and TATA boxes are thought to be transcription promoters (Dekker *et al.*, 1991; Mullineaux *et al.*, 1984; Willment *et al.*, 2007). Also present are A-tracts which are a series of A-T's that potentially play a role in regulating replication (Gutiérrez *et al.*, 1995; Suárez-López *et al.*, 1995). The key features within the SIR are the bound primer which is only found in mastrevirus (Donson *et al.*, 1984; Hayes *et al.*, 1988; Kammann *et al.*, 1991) and polyadenylation sites for the C-sense and large V-sense gene(s).

1.3.3 Movement protein (V2)

Encoded by the V2 ORF is the *mp* which is the smallest of the four genes (Fig. 1.3). Wright *et al.* (1997) showed that the transcript of this gene can sometimes be spliced. The *mp* is responsible for the movement of viral DNA from the nucleus to the cell periphery and from cell-to-cell via the plasmodesmata. Experiments using green fluorescent protein (GFP):MP constructs resulted in fluorescence present in both the original injected and adjacent cells advocating the fact that this protein is integral in cell-to-cell movement of viral DNA (Kotlizky *et al.*, 2000). Studies using *mp* gene mutants (single base deletion(s), mutation(s) or total gene replacement) were able to demonstrate that the MP is not required for replication or encapsidation of the viral DNA (Boulton *et al.*, 1989; Liu *et al.*, 1998).

It has been shown that the movement proteins of *Bean dwarf mosaic virus* (genus *Begomovirus*) can recognise DNA based on size and structural properties, and facilitate enlargement of mesophyll plasmodesmata to enable the movement of viral DNA from one cell to the next (Noueiry *et al.*, 1994; Rojas *et al.*, 1998; Sudarshana *et al.*, 1998). It is most likely that the MP of mastrevirus functions similarly, supporting this is a study by Dickinson *et al.* (1996) which shows localisation of the protein to plasmodesmata. Unlike in monopartite begomoviruses, the MP of mastreviruses has not been shown to directly interact with viral DNA, instead it most likely binds to the CP-DNA complex (Boulton, 2002; Liu *et al.*, 2001). Structural analysis of MP highlights a hydrophobic region that may interact with plant cell membrane proteins (Boulton *et al.*, 1993). The MP of MSV is implicated in symptom severity in their hosts (van der Walt *et al.*, 2008b).

1.3.4 Capsid protein (V1)

The gene which encodes the capsid protein is transcribed in the virion sense direction. This was first determined in MSV by the mapping of RNA transcripts to the MSV genome (Morris-Krsinich *et al.*, 1985). When encapsidating viral DNA this protein forms twinned icosahedral virions (Zhang *et al.*, 2001). Zhang *et al.* (2001) used both cryo-electron microscopy and *in silico* protein modelling of MSV CP to show this geminate particle has dimensions of 220 x 380 Å and is made up of 22 capsomers (each capsomer in turn is made up of five CPs). Modelling illustrated the CP structure has an eight-stranded, antiparallel β-barrel motif that is said to be common in known ssDNA viruses and an N-terminal α-helix.

Furthermore the study showed that the twinned geminate particle structure is very stable. Although no studies have looked at CP assembly dynamics of mastreviruses, a study looking at the CP of TYLCV showed they self-interacted and that the N-terminal region is needed for this interaction (Hallan & Gafni, 2001). The CP in mastreviruses has multiple functions other than simply encapsidating and protecting viral DNA. This protein also plays an essential role in accumulation of ssDNA, movement of viral DNA to the nucleus for replication (Boulton *et al.*, 1993; Liu *et al.*, 1997a; Liu *et al.*, 2001; Liu *et al.*, 1999a), and vector transmission and specificity (Boulton, 2002).

The CP of MSV binds to viral DNA (Liu *et al.*, 1997a) mediating both delivery to the nucleus and following replication of viral DNA. The CP-DNA complex interacts with the MP for movement from the nucleus to the cell periphery and for cell-to-cell movement (Kotlizky *et al.*, 2000; Lazarowitz *et al.*, 1989; Liu *et al.*, 2001; Liu *et al.*, 1999a). The CP however is not needed for viral replication (Boulton *et al.*, 1993; Boulton *et al.*, 1989).

In order for the virus to be transmitted by a vector it must be able to move through the alimentary canal into the midgut, followed by movement from the haemolymph to the salivary glands and then transmitted to the plant when the insect next feeds. The importance of the ability of the virus to cross the midgut and salivary gland barriers for transmission is evident in a study which showed MSV could be found in the head, gut and hemolymph of the known MSV vector *Cicadulina mbila* following acquisition, whereas *Digitaria streak virus* (CSMV) which cannot be transmitted by *C. mbila* was only found in the gut (Lett *et al.*, 2002). Although no one has yet shown any experimental evidence for the mechanisms behind viral movement in the vector or possible specificity determinants on the CP of mastrevirus, other studies on various geminiviruses have implicated the CP in vector specificity (Briddon *et al.*, 1990; Briddon *et al.*, 1989; Höfer *et al.*, 1997). It has been hypothesised that the CP has a recognition site that allows for receptor-mediated entry into the host epithelial cells (Lapierre & Signoret, 2004). Mastrevirus species seem to be leafhopper species specific in terms of transmission and therefore it is most likely the CP holds the key to this specificity. A study undertaken by Greber (1989) showed that although the leafhopper species *Nesoclutha pallida* was able to transmit two species of monocot-infecting mastrevirus, *Paspalum striate mosaic virus* (PSMV) and *Chloris striate mosaic virus* (CSMV), *Cicadulina bimaculata* was

unable to transmit either of these viruses. A localisation study of WDV in its leafhopper vector demonstrated that the viral particles move from the midgut through to the salivary glands and are transmitted to the plant within five minutes from initial acquisition (Wang *et al.*, 2014b). Further, an experiment using antiserum raised against WDV CP showed a reduction in accumulation of WDV throughout the insect providing evidence that the CP is central in the vector-virus interactions enabling the acquisition to transmission process (Wang *et al.*, 2014b).

1.3.5 Replication-associated protein and RepA (C1 and C1:C2)

Two replication-associated proteins known as the Rep and RepA are encoded in mastreviruses, both in the complementary sense. The Rep is produced from a splicing event which removes an intron ~85 nt (intron size varies among species) with acceptor and donor sites of GT and AG, respectively (Schalk *et al.*, 1989; Wright *et al.*, 1997). The two complementary-sense transcripts have the same start codon and share the same N-terminal sequence of ~200 amino acid (each species is slightly different), whereas the C-terminal amino acid sequence composition of each protein is different. Promoter sites for transcription of these ORFs known as TATA boxes are located within the LIR (Dekker *et al.*, 1991; Mullineaux *et al.*, 1984; Willment *et al.*, 2007). The Rep is responsible for the initiation of RCR by binding close to the origin of replication situated in the LTR and nicking the virion-sense DNA strand in the conserved nonanucleotide (TAAT[A/G]TTAC) stem loop sequence (Heyraud *et al.*, 1993). Gene modification experiments showed both Rep and RepA are essential for replication (Liu *et al.*, 1998). This study also produced a Rep mutant lacking an intron which in turn inhibited a systemic infection.

Three conserved RCR initiator motifs within the Rep are known as motif I [FLTY(P/S)], motif II [HxHxx] and motif III [YxxKx] (Ilyina & Koonin, 1992; Rosario *et al.*, 2012b) (see Fig 1.6 for overview of motifs present in representative from each mastrevirus species). The specific function of motif I is not fully understood, however, there is some evidence to suggest that it may be a Rep recognition site for iterons found in the LIR (Argüello-Astorga & Ruiz-Medrano, 2001). These authors also describe a subdomains that are associated with motif I known as iteron-related domain (IRD) and two groups of residues within this domain known as DNA-binding specificity determinants (SPD-r1 and SPD-r2). The IRD domain forms the core structure of a DNA-binding domain, consisting of a β -sheet subdomain,

empirically supported by three-dimensional protein structure analysis of a *Tomato yellow leaf curl Sardinia virus* Rep (Londoño *et al.*, 2010; Mauricio-Castillo *et al.*, 2014). A second line of evidence supports the role of IRD-motif I as an intrinsic replication specificity determinant, experimentally IRD-motif I Tomato golden mosaic virus (TGMV) mutants were shown to lose the ability to bind to DNA specifically (Orozco *et al.*, 1998).

The importance of motif II has been shown using a TYLCV model and it contains conserved histidine residues which can bind Mg^{2+} or Mn^{2+} ions and therefore is thought to play a part in metal ion coordination which is required for Rep cleavage reaction (Laufs *et al.*, 1995a). These authors also used a TYLCV model to demonstrate the importance of Motif III which is involved in cleaving the bond at the *v-ori* and subsequent joining activity of the Rep protein to the cleaved 5' end (Laufs *et al.*, 1995a; Laufs *et al.*, 1995b; Orozco & Hanley-Bowdoin, 1998)

A fourth large conserved motif known as the geminivirus Rep sequence (GRS) domain was first identified by Nash *et al.* (2011), who noted that the highest degree of conservation was at either end of the domain and therefore this may actually be two motifs with different functions. These authors investigated functionality of this domain using GRS-mutants of the begomovirus species TGMV. These mutants were non-infectious and were unable to undertake RCR and therefore they concluded this domain is intrinsic for initiation of RCR.

Downstream of motif III are three putative conserved helicase domains known as walker-A [G(P/D)(T/S)(R/S)TGK(S/T/K)(S/T/A)], walker-B [(V/I)(I/V/L)DD(I/V)] and motif C [(I/V)LxN]. These dNTP-binding domains are analogous to those in other viruses known to play a role in helicase activity (Gorbalenya & Koonin, 1993; Gorbalenya *et al.*, 1990). Amino acid sequences in this C-terminal region of WDV, MSV and CSMV also share similarities to plant transcription factor genes, known as the *myb*-like domain (Hofer *et al.*, 1992; Zhan *et al.*, 1993). Horváth *et al.* (1998) transformed yeast cells with vectors containing MSV Rep fragments and demonstrated that domains within the C-terminal can act as trans-activators in yeast cells, and therefore most likely in plant cells. Furthermore, RepA and Rep proteins have been shown to have two domains, the oligomerisation domain and N-terminal interaction domain which interact (Horváth *et al.*, 1998).

Preceding the walker-A is a conserved motif in mastreviruses with the amino acids sequence LxCxE. This motif is referred to as the retinoblastoma-related protein interaction domain (RBR interaction domain) because it has been shown to have function importance in binding the plant host retinoblastoma-related protein (Rb proteins), which would otherwise interfere with replication of the virus (Xie *et al.*, 1995). This mechanism is similar to that used by some animal viruses which have an analogous RBR interaction domain for binding the retinoblastoma protein in their animal host. RBR interaction domain mutants are unable to infect mesophyll cells of mature leaves, unlike the wild type (McGivern *et al.*, 2005). A fully intact RBR interaction domain was shown to be necessary for WDV to be able to infect wheat cells in culture (Xie *et al.*, 1995). It is worth mentioning, although a RBR interaction domain amino acid sequence of LxCxE is conserved in all dicot-infecting mastreviruses, MSV and WDV, it does however vary among other species of mastrevirus (see Fig. 1.6 and Additional fig. 1.1 for overview). Only a few amino acids downstream of this motif is a newly identified motif known as the RxL motif. Mutations in the RxL motif of the begomovirus *African cassava mosaic virus* (ACMV) Rep have been shown to render the virus unable to cause a symptomatic infection in tobacco plants. In yeast this mutant inhibited re-replication following introduction of a wildtype ACMV Rep (Hipp *et al.*, 2014). This evidence implicates the RxL motif as an interaction site for a novel mechanism linking the Rep to host proteins involved in the cell cycle.

The RepA protein in particular has been shown to bind to Rb proteins unlike the Rep (Collin *et al.*, 1996; Liu *et al.*, 1999b). The C-terminal region that is unique to RepA has been experimentally shown, using yeast cells, to interact with a member from a specific group of host factor proteins known as Geminivirus Rep A-binding (GRAB) proteins with this region is referred to as the GRAB interaction domain (Xie *et al.*, 1999). Small alterations in the amino acid sequence of this motif in *Chickpea chlorotic dwarf virus* (CpCDV; formerly known Bean yellow dwarf virus (BeYDV), and MSV resulted in the retention of virus infectivity and symptom induction (although at a reduced efficiency in planta (Liu *et al.*, 1999b; Shepherd *et al.*, 2005). Collin *et al.* (1996) demonstrated that RepA is required for V-sense gene expression (trans-activation domain) and that an interaction between RepA and Rb proteins may be essential for expression. The *rep* of WDV has been shown to inhibit both local and systemic RNA silencing by binding small interfering siRNA's in transgenic *Nicotiana benthamiana* (Wang *et al.*, 2014a).

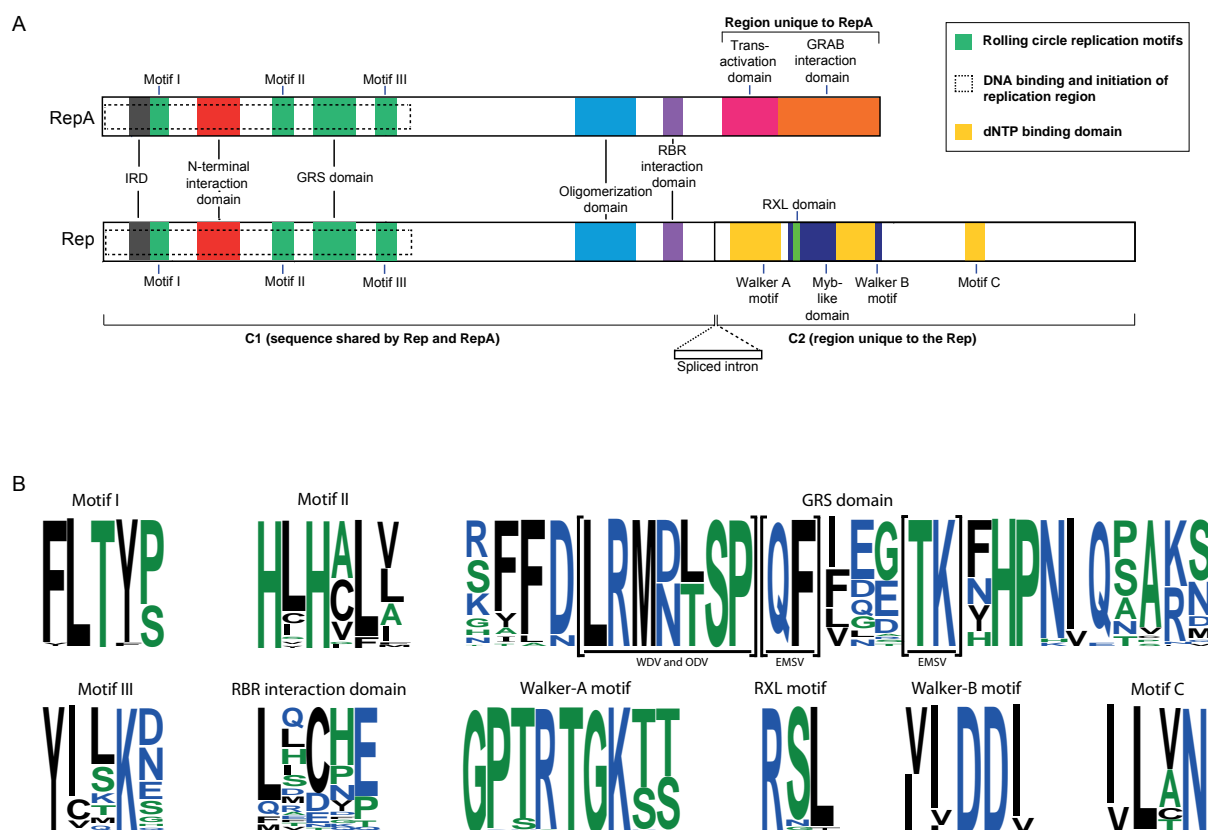


Figure 1.6 Summary of the functional motifs identified in the Rep and RepA of representatives of each mastrevirus species. **A)** The Rep and RepA share the N-terminal sequence whereas the C-terminal for each is unique. In the shared N-terminal of both Reps are three motifs (I, II and III) which are rolling circle replication motifs (Argüello-Astorga & Ruiz-Medrano, 2001; Gutierrez, 1999; Laufs *et al.*, 1995b; Nash *et al.*, 2011; Orozco & Hanley-Bowdoin, 1998). An iteron-related domain which likely binds to iterons in the LIR precedes motif I (Argüello-Astorga & Ruiz-Medrano, 2001). Also found in both proteins are two domains which interact, the N-terminal interaction domain and the oligomerisation domain (Horváth *et al.*, 1998), and the RBR interaction domain which interacts with the host retinoblastoma-related proteins (Xie *et al.*, 1995). Unique to RepA is the trans-activation domain which is involved in activating transcription of the virion-sense genes (Collin *et al.*, 1996), and a GRAB interaction domain which interacts with host factors (Xie *et al.*, 1999). Those motifs which are unique to the Rep, located in the C-terminal region are walker-A, walker-B and motif-C, are dNTP binding motifs, thought to be involved in the Rep helicase activity (Gorbalenya & Koonin, 1993; Gorbalenya *et al.*, 1990). The RXL motif is a possible cyclin-interacting motif. Also in the C-terminal region is the *myb*-like domain which is also thought to interact with cellular proteins (Hofer *et al.*, 1992). **B)** To show the level of conservation of each motif amino acid sequence within the Rep of mastreviruses, representatives from each mastrevirus species were aligned and using amino acid sequence logo the degree of which a specific amino acid is present at each position is shown by the relative size of each letter. The GRS domain of three species, WDV, ODV and EMSV contain sequence insertions not found in any other species, these are shown by brackets and the species this insertion is found in is noted underneath each bracketed sequence. Colour of amino acid designates hydrophobicity; blue is hydrophobic, green is neutral and black hydrophobic.

1.3.6 Monocot-infecting mastreviruses

The majority of mastrevirus infect members of the *Poaceae* family, to date twenty-five species have been documented, all from the old world (Fig. 1.2; Table 1.2). Twelve species, have been found in Africa and the south-west Indian Ocean islands (SWIO) and nine in Australia. Three species have been documented in Eurasia. One species has been identified in Japan and one in Vanuatu. Symptoms induced by monocot-infecting mastrevirus are similar across hosts. These symptoms are chlorosis in the form of striations and/or mosaic patterns, however, symptoms in maize and cereal crops can be more severe resulting in reduction of crop yield and sometimes cause stunting of plant.

1.3.6.1 African streak mastreviruses

Of the twelve species identified in Africa, MSV has been the most extensively studied. This is largely due to the devastating impact this virus has on maize, a staple crop throughout Africa. Maize was introduced to the continent of Africa in the 16th century where it became a staple crop, with the first description of the effects of MSV on crops was by Storey (1925). MSV infects not only maize but sugarcane and a wide range of wild grass species. A total of eleven strains of MSV have been described MSV-A to MSV-K. All eleven strains have been found infecting a wide range of grass species, however, MSV-A is the only strain which has been identified to infect maize in the field (Monjane *et al.*, 2011a; Oluwafemi *et al.*, 2011; Shepherd *et al.*, 2008a; Shepherd *et al.*, 2010; Varsani *et al.*, 2008b). Strains MSV-B – -K are all adapted to infect wild-grass species in the genera *Digitaria*, *Urochloa* and *Setaria*. Many recombination events among the different strains are evident which has led to the diversification of MSV (Monjane *et al.*, 2011a; Owor *et al.*, 2007; Varsani *et al.*, 2009a). An recombination event which occurred between ancestral MSV-B and MSV-G/-F variants is attributed to have resulted in the emergence of the maize-adapted MSV-A (Varsani *et al.*, 2008b). Dating analysis of this events indicates it most likely occurred mid-19th century, ~20 years prior to the first documented reports of maize streak disease in South Africa (Harkins *et al.*, 2009b). Subsequent movement of MSV-A subtypes throughout the African continent has been reconstructed with various subtypes identified in distinct geographical regions (Monjane *et al.*, 2011b). MSV is vectored by the leafhopper genus *Cicadulina* (Bosque-Pérez, 2000), with nine species within this genus being known to transmit MSV (Bigirwa *et al.*, 1995; Dabrowski, 1987; Okoth *et al.*, 1987; Rose, 1962; Storey, 1924; 1936; Webb, 1987). *Maize streak Reunion virus* (MSRV) is the only other known mastrevirus species to infect maize. It

was first identified in maize plants on the island of La Réunion (Pande *et al.*, 2012) and has subsequently been found in *Setaria barbata* and *Rottboellia* sp. from Nigeria (Oluwafemi *et al.*, 2014).

Five species from Africa which are known only to infect wild grass species are *Panicum streak virus* (PanSV), *Axonopus compressus streak virus* (ACSV), *Eragrostis minor streak virus* (EMSV), *Eragrostic streak virus* (ESV) and *Urochloa streak virus* (USV). Of these species PanSV has been the most extensively sampled with nine characterised strains, this species has a similar level of diversity to that seen in MSV and is known to infect five wild grass species. It is vectored by one of the same leafhopper species as MSV, *C. mbila* (Briddon *et al.*, 1992). ACSV is a newly described species from Nigeria isolated from *A. compressus* (Oluwafemi *et al.*, 2014). Two species EMSV and ESV both infect *Eragrostis* sp. and USV infects *Urochloa deflexa*.

SacSV, SacSV, SWSV, SSRV, SSV are collectively known as the sugarcane-infecting streak viruses, these viruses are all from Africa and the SWIO and have predominantly been isolated from sugarcane. Two species, SSV and SSRV have also been isolated from wild grasses. Two strains have been identified in the species SSV and SSRV, and three in the newly discovered species, SWSV (Candresse *et al.*, 2014).

1.3.6.2 Australian striate mosaic mastreviruses

All species of monocot-infecting mastrevirus from Australia have been recovered from wild grasses. *Bromus catharticus striate mosaicvirus* (BCSMV), *Digitaria ciliaris striate mosaic virus* (DCSMV), *Digitaria didactyla striate mosaic virus* (DDSMV), *Sporobolus striate mosaic virus* (SSMV) -1 and -2 have each only ever been recovered from a single *Poaceae* sp., PDSMV, *Paspalum striate mosaic virus* (PSMV) and *Chloris striate mosaic virus* (CSMV) infect a wide range of *Poaceae* sp. (Kraberger *et al.*, 2012) (see Table 1 in Chapter Two for a full list of host species). PSMV and DCSMV both have two designated strains, PSMV-A and B, and DCSMV-A and B, respectively. The leaf hopper vector *Neoclutha pallida* is known to transmit PSMV, BCSMV, DDSMV and CSMV with varying efficiency. This information is based on a single study undertaken by Greber (1989) and no other

investigations have been undertaken into possible vectors of Australian monocot-infecting mastrevirus.

1.3.6.3 Japan-Pacific mastreviruses

Both *Miscanthus streak virus* (MiSV) from Japan and *Digitaria streak virus* (DSV) from Vanuatu have single representative genomes deposited in GenBank with each being named after their single known hosts, *Miscanthus sacchariflorus* and *Digitaria sanguinalis*, respectively. No vector is yet known for either virus.

1.3.6.4 Eurasian mastreviruses

WDV, WDIV and ODV are important pathogens of wheat, barley and oat in Europe and Asia. The most economically damaging of the cereal-infecting mastrevirus is WDV, of which there are five strains, WDV-A to WDV-E which are known to infect wheat (*T. aestivum*), barley (*Hordeum vulgare*), oat (*Avena sativa*) and two wild grass species *Lolium* sp., *Secale* sp. and are transmitted by the leafhopper species *Psammotettix alienus* and *Psammotettix provincialis* (Ekzayez *et al.*, 2011; Schubert *et al.*, 2007; Schubert *et al.*, 2013; Wang *et al.*, 2014b). WDIV has only been isolated from wheat (*Triticum aestivum*) in India and has recently been associated with a satellite molecules (Kumar *et al.*, 2014; Kumar *et al.*, 2012). ODV is also vectored by *P. alienus* and has only been documented in Germany infecting oats (Schubert *et al.*, 2007).

1.3.7 Dicot-infecting mastreviruses

There are six species of mastreviruses which infect dicotyledonous plants. One species is found outside of Australia, *Chickpea chlorotic dwarf virus* (CpCDV) (Horn *et al.*, 1993; Krabberger *et al.*, 2013a; Liu *et al.*, 1997b; Nahid *et al.*, 2008) which is found in South Africa, North-east Africa, the Middle East, Turkey and the Indian subcontinent. The other five species, *Chickpea chlorosis virus* (CpCV) (Hadfield *et al.*, 2012; Krabberger *et al.*, 2013a; Thomas *et al.*, 2010), *Chickpea chlorosis Australia virus* (CpCAV) (Hadfield *et al.*, 2011) and *Tobacco yellow dwarf virus* (TYDV) (Hadfield *et al.*, 2012; Morris *et al.*, 1992), *Chickpea redleaf virus* (CpRLV) (Thomas *et al.*, 2010), *Chickpea yellows virus* (CpYV) (Hadfield *et al.*, 2012) have all only been found in Australia (Fig. 1.2 and Table 1.2).

CpCDV, through molecular techniques, has thus far been identified in chickpeas (*Cicer arietinum*) (Horn *et al.*, 1993; Krabberger *et al.*, 2013a; Mumtaz *et al.*, 2011; Nahid *et al.*,

2008), lentils (*Lens culinaris*) (Kraberger *et al.*, 2013a), faba bean (*Vicia faba*) (Kraberger *et al.*, 2013a), french bean (*Phaseolus* sp.) (Halley-Stott *et al.*, 2007; Liu *et al.*, 1997b), field pea (*Pisum sativum*) (Kraberger *et al.*, 2013a), sugarbeet (*Beta vulgaris*) (Farzadfar *et al.*, 2008), the weed host *Sesbenia bispinosa* (Nahid *et al.*, 2008) and most recently capsicum (*Capsicum annuum*) (Akhtar *et al.*, 2013), and cotton (Manzoor *et al.*, 2014). CpCDV in sugar beet has only been identified by partial genome sequencing. Symptoms in legumes can include stunting, chlorosis and/or reddening of the leaves. Sugar beet presents mild chlorosis and stunting and capsicum upwards leaf cupping and stunting. CpCDV has only been isolated from cotton with a mixed infection (begomovirus; *Cotton leaf curl Burewala virus*) therefore symptoms caused by CpCDV alone are unclear. There are twelve characterised strains of CpCDV, CpCDV-A–CpCDV-L. CpCDV is known to be vectored by the leafhopper species *Orosius orientalis* (Horn *et al.*, 1994; Horn *et al.*, 1993) and *Orosius albicinctus* (Akhtar *et al.*, 2011).

Four species of Australian dicot-infecting mastreviruses, CpCV, CpCAV, CpRLV and CpYV are all known to infect chickpeas (Hadfield *et al.*, 2012; Kraberger *et al.*, 2013a; Schwinghamer *et al.*, 2010; Thomas *et al.*, 2010), the first two species have also been isolated from French bean (Hadfield *et al.*, 2012). Classical symptoms in chickpeas include chlorosis or reddening of leaves, stunting and often browning of the phloem. Five strains of CpCV are currently known, CpCV-A–CpCV-F. No vector is known for CpCV, CpCAV, CpRLV or CpYV. TYDV has been isolated from tobacco, french bean and chickpea. A partial Rep sequence has been identified from turnip weed (*Rapistrum rugosum*) which potentially is a distinct strain of TYDV based on nucleotide pairwise identity comparison of this partial sequence with other known TYDV sequences (Schwinghamer *et al.*, 2010; Thomas *et al.*, 2010). Symptoms in tobacco include down-curling of leaves, chlorosis, stunting and sometimes necrotic areas in ageing leaves. The leafhopper vector *O. orientalis* which is known to vector CpCDV, is also present in Australia (Fletcher, 2009). A study sampled leafhoppers from the species *Orosius orientalis* and *Anzygina zealandica* in south-east Australia which tested positive for TYDV (Trębicki *et al.*, 2010), however, no transmission experiments have been done to confirm either of these species are able to transmit TYDV.

Table 1.2 Details of representative mastrevirus species and strains for which full genomes have been recovered. Accession numbers, host species, country and associated reference are included. Adapted from Muhire *et al.* (2013)

Species	Strain [GenBank no.]	Host	Country	References
<i>Axonopus compressus streak virus</i>	ACSV [KJ437671]	<i>Axonopus compressus</i>	Nigeria	(Oluwafemi <i>et al.</i> , 2014)
<i>Bromus catharticus striate mosaic virus</i>	BCSMV [HQ113104]	<i>Bromus catharticus</i>	Australia	(Gafni & Epel, 2002; Greber, 1989; Hadfield <i>et al.</i> , 2011)
<i>Chickpea Australia virus</i>	CpCAV [JN989420]	<i>Cicer arietinum</i> <i>Phaseolus</i> sp.	Australia	(Hadfield <i>et al.</i> , 2012)
<i>Chickpea chlorotic dwarf virus</i>	CpCDV-A [FR687959]	<i>Cicer arietinum</i> , <i>Pisum sativum</i>	Syria, Turkey, Iran	(Akhtar <i>et al.</i> , 2013; Ali <i>et al.</i> , 2004; Halley-Stott <i>et al.</i> , 2007; Krabberger <i>et al.</i> , 2013a; Manzoor <i>et al.</i> , 2014; Mumtaz <i>et al.</i> , 2011; Nahid <i>et al.</i> , 2008)
	CpCDV-B [Y11023]	<i>Cicer arietinum</i> , <i>Phaseolus vulgaris</i>	Pakistan, South Africa	
	CpCDV-C [AM849097]	<i>Cicer arietinum</i>	Pakistan	
	CpCDV-D [FR687960]	<i>Cicer arietinum</i> , <i>Pisum sativum</i>	Pakistan, India	
	CpCDV-E [AM933135]	<i>Cicer arietinum</i>	Sudan	
	CpCDV-F [KC172666]	<i>Cicer arietinum</i> , <i>Capsicum annuum</i>	Pakistan, Yemen, Syria, Oman	
	CpCDV-G [KC172674]	<i>Cicer arietinum</i>	Eritrea	
	CpCDV-H [KC172676]	<i>Cicer arietinum</i>	Eritrea	
	CpCDV-I [KC172677]	<i>Cicer arietinum</i>	Eritrea	
	CpCDV-J [KC172678]	<i>Cicer arietinum</i>	Eritrea	
	CpCDV-K [KM229905]	<i>Cicer arietinum</i>	Eritrea	
	CpCDV-L [HE864164]	<i>Gossypium hirsutum</i>	Pakistan	
<i>Chickpea chlorosis virus</i>	CpCV-A [JN989415]	<i>Cicer arietinum</i>	Australia	(Hadfield <i>et al.</i> , 2012; Krabberger <i>et al.</i> , 2013a; Thomas <i>et al.</i> , 2010)
	CpCV-B [GU256531]	<i>Cicer arietinum</i>	Australia	
	CpCV-C [JN989416]	<i>Cicer arietinum</i>	Australia	
	CpCV-E [JN989426]	<i>Cicer arietinum</i> , <i>Phaseolus</i> sp.	Australia	
<i>Chickpea redleaf virus</i>	CpRLV [GU256532]	<i>Cicer arietinum</i>	Australia	(Thomas <i>et al.</i> , 2010)
<i>Chickpea yellows virus</i>	CpYV [JN989439]	<i>Cicer arietinum</i>	Australia	(Hadfield <i>et al.</i> , 2012)
<i>Chloris striate mosaic virus</i>	CSMV [M20021]	<i>Chloris gayana</i> , <i>Eriochloa polystachya</i> , <i>Paspalum dilatatum</i> , <i>Triticum aestivum</i> , <i>Panicum</i> sp., <i>Sporobolus</i> sp., <i>Digitaria ciliaris</i>	Australia	(Andersen <i>et al.</i> , 1988; Greber, 1989)
<i>Digitaria ciliaris striate mosaic virus</i>	DCSMV-A [JQ948091]	<i>Digitaria ciliaris</i>	Australia	(Krabberger <i>et al.</i> , 2012)
	DCSMV-B [JQ948088]	<i>Digitaria ciliaris</i>	Australia	
<i>Digitaria didactyla striate mosaic virus</i>	DDSMV [HM122238]	<i>Digitaria didactyla</i>	Australia	(Briddon <i>et al.</i> , 2010b; Greber, 1989)
<i>Digitaria streak virus</i>	DSV [M23022]	<i>Digitaria sanguinalis</i>	Vanuatu	(Donson <i>et al.</i> , 1987)
<i>Eragrostis minor streak virus</i>	EMSV [JF508490]	<i>Eragrostis minor</i>	Namibia	(Martin <i>et al.</i> , 2011c)
<i>Eragrostis streak virus</i>	ESV [EU244915]	<i>Eragrostis curvula</i>	Zimbabwe	(Shepherd <i>et al.</i> , 2008b)

Species	Strain [GenBank no.]	Host	Country	References
<i>Miscanthus streak virus</i>	MiSV [E02258]	<i>Miscanthus sacchariflorus</i>	Japan	(Chatani <i>et al.</i> , 1991)
<i>Maize streak reunion virus</i>	MSRV [JQ624879]	<i>Zea mays</i> , <i>Setaria barbata</i> , <i>Rottboellia</i> sp	La Reunion and Nigeria	(Oluwafemi <i>et al.</i> , 2014; Pande <i>et al.</i> , 2012)
<i>Maize streak virus</i>	MSV-A [Y00514]	<i>Zea mays</i> <i>Axonopus compressus</i> <i>Cenchrus myosuroides</i> <i>Digitaria</i> sp. <i>Eragrostis curvula</i> <i>Ehrharta calycina</i> <i>Eustachys petraea</i> <i>Pennisetum</i> sp. <i>Rattara petiolata</i> <i>Rottboellia cochinchinensis</i> , <i>Saccharum</i> sp. <i>Setaria</i> sp. <i>Saccharum</i> sp. <i>Urochloa maxima</i>	Burkina Faso, Cameroon, Central African Republic, Chad, Kenya, La Reunion, Lesotho, Mozambique, Nigeria, South Africa, Uganda, Zambia, Zimbabwe	(Harkins <i>et al.</i> , 2009b; Martin <i>et al.</i> , 2001; Monjane <i>et al.</i> , 2011b; Owor <i>et al.</i> , 2007; Pande <i>et al.</i> , 2012; Shepherd <i>et al.</i> , 2010; Shepherd <i>et al.</i> , 2008b)
	MSV-B [EU628597]	<i>Avena sativa</i> <i>Cenchrus myosuroides</i> <i>Digitaria</i> sp. <i>Ehrharta calycina</i> <i>Hordeum vulgare</i> <i>Lolium rigidum</i> <i>Rattara petiolata</i> <i>Setaria grisebachii</i> <i>Urochloa maxima</i> <i>Urochloa plantaginea</i>	La Reunion, Uganda, Rwanda, Kenya, South Africa, Mozambique	
	MSV-C [AF007881]	<i>Setaria</i> sp.	South Africa, Uganda	
	MSV-D [AF329889]	<i>Urochloa</i> sp.	South Africa	
	MSV-E [EU628626]	<i>Digitaria ciliaris</i> <i>Setaria barbata</i>	Mozambique, South Africa, Uganda	
	MSV-F [EU628629]	<i>Urochloa maxima</i> <i>Digitaria ciliaris</i>	Burundi, Uganda, Nigeria	
	MSV-G [EU628631]	<i>Brachiaria deflexa</i> <i>Brachiaria lata</i> <i>Digitaria</i> sp. <i>Panicum maximum</i> <i>Paspalum notatum</i>	Nigeria, Mali	
	MSV-H [EU628638]	<i>Setaria barbata</i>	Nigeria	
	MSV-I [EU628639]	<i>Digitaria ciliaris</i>	South Africa	
	MSV-J [EU628641]	<i>Pennisetum</i> sp	Zimbabwe	
	MSV-K [EU628643]	<i>Eustachys petraea</i> , <i>Setaria verticillata</i>	Uganda, Zimbabwe	
<i>Oat dwarf virus</i>	ODV-[AM296025]	<i>Avena sativa</i>	Germany	(Schubert <i>et al.</i> , 2007)

Species	Strain [GenBank no.]	Host	Country	References
<i>Panicum streak virus</i>	PanSV-A [L39638]	<i>Ehrharta calycina</i>	Zimbabwe, South Africa,	(Rybicki, 1994; Varsani <i>et al.</i> , 2009a; Varsani <i>et al.</i> , 2008a)
	PanSV-B [X60168]	<i>Panicum maximum</i>	Mozambique	
	PanSV-C [EU224264]	<i>Panicum maximum</i>	Kenya	
	PanSV-D [EU224265]	<i>Urochloa plantaginea</i>	Zimbabwe	
	PanSV-E [GQ415389]	<i>Urochloa maxima</i>	Nigeria	
		<i>Panicum maximum</i>	Kenya	
	PanSV-F [GQ415392]	<i>Panicum maximum</i>	Kenya	
	PanSV-G [GQ415396]	<i>Panicum maximum</i>	Mayotte	
<i>Paspalum dilatatum striate mosaic virus</i>	PanSV-H [GQ415397]	<i>Panicum maximum, Brachiaria deflexa</i>	Nigeria, Central African Republic	(Kraberger <i>et al.</i> , 2012)
	PanSV-I [GQ415401]	<i>Panicum tricholadum</i> <i>Brachiaria deflexa</i>	Kenya	
	PDSMV [JQ948087]	<i>Paspalum dilatatum</i>	Australia	
		<i>Digitaria ciliaris</i>		
	PSMV-A [JF905486]	<i>Paspalum dilatatum</i>	Australia	
		<i>Digitaria ciliaris</i>		
		<i>Ehrharta erecta</i>		
	PSMV-B [JQ948069]	<i>Paspalum dilatatum</i>	Australia	
<i>Saccharum streak virus</i>	SacSV [GQ273988]	<i>Saccharum</i> sp.	South Africa	(Lawry <i>et al.</i> , 2009)
<i>Sugarcane streak Egypt virus</i>	SSEV [AF239159]	<i>Saccharum</i> sp.	Egypt	(Bigarré <i>et al.</i> , 1999)
<i>Sugarcane white streak virus</i>	SWSV-A [KJ187746]	<i>Saccharum</i> sp.	Egypt	
	SWSV-B [KJ187747]	<i>Saccharum</i> sp.	Sudan	
	SWSV-C [KJ187749]	<i>Saccharum</i> sp.	Sudan	
<i>Sporobolus striate mosaic virus 1</i>	SSMV 1 [JQ948051]	<i>Sporobolus australasicus</i>	Australia	(Kraberger <i>et al.</i> , 2012)
<i>Sporobolus striate mosaic virus 2</i>	SSMV 2 [JQ948052]	<i>Sporobolus australasicus</i>	Australia	(Kraberger <i>et al.</i> , 2012)
<i>Sugarcane streak reunion virus</i>	SSRV-A [AF072672]	<i>Saccharum</i> sp., <i>Setaria barbata</i>	La Reunion	(Bigarré <i>et al.</i> , 1999; Shepherd <i>et al.</i> , 2008b)
	SSRV-B [EU244916]	<i>Paspalum conjugatum</i>	Zimbabwe	
<i>Sugarcane streak virus</i>	SSV-A [M82918]	<i>Saccharum</i>	South Africa	(Hughes <i>et al.</i> , 1993; Shepherd <i>et al.</i> , 2008b)
	SSV-B [EU244914]	<i>Cenchrus myosuroides</i>	La Reunion	
<i>Tobacco yellow dwarf virus</i>	TYDV-A [M81103]	<i>Nicotiana</i> sp., <i>Phaseolus</i> sp., <i>Cicer arietinum</i>	Australia	(Hadfield <i>et al.</i> , 2012; Kraberger <i>et al.</i> , 2013a; Morris <i>et al.</i> , 1992)
<i>Urochloa streak virus</i>	USV-[EU445697]	<i>Urochloa deflexa</i>	Nigeria	(Oluwafemi <i>et al.</i> , 2014; Oluwafemi <i>et al.</i> , 2008)
<i>Wheat dwarf India virus</i>	WDIV [JQ361910]	<i>Triticum aestivum</i>	India	(Kumar <i>et al.</i> , 2012)
<i>Wheat dwarf virus</i>	WDV-A [AJ783960]	<i>Hordeum vulgare</i> , <i>Avena sativa</i>	Bulgaria, Czech Republic, Germany, Hungary, Turkey, Ukraine	(Kacprzak <i>et al.</i> , 2005; Köklü <i>et al.</i> , 2007; Kvarnheden <i>et al.</i> , 2002; MacDowell <i>et al.</i> , 1985; Tobias <i>et al.</i> , 2011)
	WDV-B [FJ620684]	<i>Hordeum vulgare</i>	Iran	
	WDV-C [JQ647455]	<i>Triticum aestivum</i>	China, Hungary, Tibet	
	WDV-D [JN791096]	<i>Hordeum vulgare</i>	Iran	
	WDV-E [AM040732]	<i>Triticum aestivum</i>	China, Czech Republic, Hungary, France, Germany, Iran, Sweden, Ukraine	
		<i>Lolium</i> sp.		
		<i>Secale</i> sp.		

1.4 Mastrevirus detection methods

1.4.1 Serology

Serological assays have long been a common method for diagnostic testing of samples from the field (Cenchrus & Coix, 1991; Greber, 1989; Kumari *et al.*, 2006; Kumari *et al.*, 2004; Kumari *et al.*, 2008; Makkouk *et al.*, 2003a; Makkouk *et al.*, 2003b; Thomas *et al.*, 2010). Polyclonal antiserum has typically been used for different serological tests such as double-antibody sandwich enzyme-linked immunosorbent assays (ELISA), direct antigen-coating ELISA and dot-blot ELISA (Greber, 1989; Kumari *et al.*, 2006; Liu *et al.*, 1997b; Thomas *et al.*, 2010). MSV monoclonal antibodies have also been used for identification of serologically different MSV isolates (Cenchrus & Coix, 1991). Although serological assays can be a rapid and an effective way to perform diagnostic testing, some assays have been shown to cross react. For example anti bodies raised against CpCDV can also cross react with Australian dicot-infecting mastreviruses, and therefore cannot always give an unequivocal diagnosis (Horn *et al.*, 1993; Liu *et al.*, 1997b). Due to the cross reactivity this is also not an effective method for identification of different species of mastrevirus.

1.4.2 Polymerase chain reaction

An early method for investigating mastrevirus diversity was PCR amplification using degenerative primers which are able to be used on a range of genotypes within closely related mastreviruses followed by restriction fragment length polymorphism (RFLP) and/or partial genome sequencing (Martin *et al.*, 2001; Willment *et al.*, 2001). RFLP, to certain extent can be used to determine variants of various species and strains based on the restriction patterns. This method can be time consuming and is not ideal for classification of new variants in the current environment where sequencing is relative cheap.

1.4.3 Rolling circle amplification

Phi29 is a proof reading polymerase with 3'-5' exonuclease activity, this enzyme is derived from a bacteriophage which infects the bacteria *Bacillus subtilis*. This polymerase binds strongly to ssDNA and has a strand displacement mechanism enabling the replication of circular DNA (Nelson *et al.*, 2002). Phi29 polymerase in conjunction with random hexamers

enables the amplification of low levels of viral DNA without any prior knowledge of template sequence. This has revolutionised the molecular approaches to geminivirus discovery and molecular experimental approach to circular DNA viral research in general (Haible *et al.*, 2006; Inoue-Nagata *et al.*, 2004; Niel *et al.*, 2005; Shepherd *et al.*, 2008a). This method has been the foundation for the rapid increase in our knowledge of mastrevirus diversity (Harkins *et al.*, 2009b; Owor *et al.*, 2007; Shepherd *et al.*, 2008a; Shepherd *et al.*, 2010; Varsani *et al.*, 2009b; Varsani *et al.*, 2008b) and is often used to enrich circular molecules prior to performing next NGS on a sample for the discovery of novel ssDNA viruses, plasmids and circular genetic elements (Jørgensen *et al.*, 2014; Labonté & Suttle, 2013; Ng *et al.*, 2011a; Roux *et al.*, 2012; Sikorski *et al.*, 2013; Zawar-Reza *et al.*, 2014)

1.5 Next-Generation sequencing (NGS)

Advances in sequencing technology over the last decade have hugely expanded the scope and capabilities of what scientists can now achieve in a shorter time and for a fraction of the cost. High-throughput sequencing, also known as next-generation sequencing allows scientists not only to sequence a large number of reads in a short time, but non-specific amplification means no prior knowledge of target DNA sequences is necessary (Metzker, 2010). Several NGS platforms exist, each with their own advantages and disadvantages. Template preparation regardless of platform generally begins with the shearing of purified DNA sequences into smaller DNA fragments and in turn these fragments are ligated to oligo adapter(s) with the whole process, being referred to as preparation of a DNA library (Shendure & Ji, 2008). The DNA adapters are used to tether the fragmented DNA to a surface in order to immobilise it for the next stage. Although the chemistry and mechanisms behind the next steps vary between platforms, the central chemistry of most platforms involves nucleotides which are dye labelled and have a terminating functional group which is reversible (Metzker, 2010). There are several NGS platforms, however, the three platforms most universally used are Roche/454's GS FLX system, Illumina/Solexa's GA HiSeq system and Applied Biosystems/SOLiD system. The following is a brief overview of the processes and chemistry of these three platforms.

1.5.1 Roche/454's GS FLX system

A library of small DNA fragments ligated to oligo adapters are resuspended into an oil emulsion where they are bound to DNA-capture beads that have complementary sequences to

the oligo adapters on their surface. Each oil immersed capsule contains a DNA template for sequencing which is bound to a bead and reagents for amplification by PCR. Individual oil emulsion capsules containing a single bead are put through PCR temperature cycling to amplify each template. Emulsion is broken and the untethered complementary strands are washed away leaving the beads enriched. These beads are incubated with polymerase and then along with a solution containing single-stranded binding proteins and are then added to a 454 picotiter plate where each bead is fixed to the plate in an individual well. Smaller beads with bound active enzymes needed for the biochemical reactions (containing ATP sulfurylase and luciferase, required for pyrosequencing) are added to the plate. In each well a single nucleotide is added at a time and if incorporated the pyrophosphate is freed when ATP sulfurylase and luciferase are incorporated and a light signal is generated. This light signal is registered live in each well by a fibre-optic bundle. Reads produced by 454 pyrosequencing generally produces reads of ~600 bp, with a maximum of ~1000 bp, which are the longest reads out of these three NGS systems. The major downside when using this system is when the machine is measuring a stretch of the same base incorporations in a row signal strength can be difficult to gauge and therefore is prone to insertion or deletion errors (Mardis, 2008; Shendure & Ji, 2008; Zhang *et al.*, 2011).

1.5.2 Illumina/Solexa's GA HiSeq system

Library prepared template DNA fragments of approximately 300 bp in size have oligo adapters ligated to each end. One of the adapters is then bound to linker that is fixed to a surface made up of channels so that each template will remain fixed in a cluster throughout process. Bridge PCR then takes place and denaturation using formamide occurs after each cycle, resulting in several 1,000 copies from an individual (single-stranded) template are produced in a cluster. Universal primers are added which bind to adapters preceding the template. A single reversible terminator base is then added per cycle, the fluorescent label is cleaved off and a burst of light unique to each of the four bases is captured. This system has high raw base accuracy of >99.5 %, resulting in single pair-end reads ranging from 2 x 100 bp to 2 x 300 bp, dependent on which model of machine. The resulting reads are much smaller than those from Roche 454 sequencing (Metzker, 2010; Shendure & Ji, 2008; Zhang *et al.*, 2011).

1.5.3 Biosystems/SOLiD system

Following library preparation, a SOLiD system sequencing workflow is as follows. Template DNA fragments, PCR reaction components, beads and primers are combined in an oil emulsion to produce clonal populations of each DNA template attached to a single bead. Following the PCR, each emulsion is dissolved and beads with clonally amplified DNA fragments are recovered. Each bead is then covalently fixed to the surface of a glass slide and a universal primer which is complementary to the adapter is added. The primer hybridises to adapters on each bead and is followed by the addition of DNA ligase and four fluorescently labelled di-based probes (octamer oligonucleotides), each labelled with a unique fluorescent dye. There are 16 possible di-nucleotides sequences containing different two base combinations, four of these are added at a time. The probes complementary to the template hybridise to it and are ligated to the universal primer. Following ligation the probe fluoresces and this is captured for all templates using an imaging system. The dye is then cleaved off leaving a 5' phosphate group open for the next probe. This process is repeated for several cycles, resulting in every fifth base being sequenced. The universal primer and synthesised strand are then removed through washing steps and a second round (referred to as primer reset) of this process is initiated by the addition of another universal primer which binds one position in front of the last universal primer for sequencing of the next frame as well as probes and DNA ligase. Primer reset is repeated five times in total. This method of sequencing allows for high quality sequencing results because each base is checked twice in separate cycles. Reads can be between 60 and 75 bp in length, a lot shorter than with other systems which can make them hard to assemble without a scaffold (Mardis, 2008; Shendure & Ji, 2008).

1.6 NGS approaches for the discovery of novel geminiviruses and other Rep encoding ssDNA viruses

In recent years NGS has proven to be a useful tool in the discovery of novel geminiviruses, some of which are highly divergent. Emerging geminiviruses can pose a serious threat to agriculture and prior to the availability of NGS the surveillance of these viruses was biased to those species that were known. Several studies have shown that NGS can be used to identify novel geminiviral genomes using viral DNA or short interfering RNA (siRNA) analyses

methods (Kreuze *et al.*, 2009; Loconsole *et al.*, 2012; Massart *et al.*, 2014; Seguin *et al.*, 2014). The initial process of sample preparation of viral DNA and siRNA templates differs; purified viral DNA is enriched using bacteriophage Phi29 DNA polymerase (a high fidelity enzyme which preferentially amplifies circular DNA through strand displacement) and random hexamer primers whereas siRNA is first converted to cDNA. Following these two processes, templates can be library prepped for the desired NGS platform. Illumina/Solex and Roche 454 systems are the most commonly used platforms for the discovery of novel viruses. Examples of newly discovered geminivirus using a NGS informed approaches are the highly divergent geminiviruses, CCDAV (Loconsole *et al.*, 2012) and GCFaV (Poojari *et al.*, 2013) and a new sugarcane infecting mastrevirus species, SWSV, of which three new strains were identified (Candresse *et al.*, 2014). Kreuze *et al.* (2009) was one of the pioneering studies for the use of small RNAs as a template in NGS identification of mastrevirus-like sequences in sweet potato (Kreuze *et al.*, 2009), however, a full genome was never recovered.

NGS has also been used as a tool for the discovery of novel circular ssDNA viruses. A metagenomic approach allows for sequencing of all nucleic acid molecules within a sample. It is a combination of traditional molecular techniques, rolling circle amplification enrichment with bacteriophage Phi29 DNA polymerase and NGS that is now commonly used for identification of novel circular ssDNA viruses. This is even more the case as NGS sequencing becomes significantly affordable, especially if samples are multiplexed.

Viral metagenomic approaches using NGS have led to the discovery of novel circular ssDNA viral genomes that share some similarity to the known families *Geminiviridae*, *Nanoviridae*, *Circoviridae*, and the recently proposed genera cyclovirus within the *Circoviridae* family and a potential new family gemycircularvirus. These novel viruses in most cases only share similarities to known viruses based on common motifs found in the Rep and these novel viruses are currently referred to as circular Rep-encoding ssDNA (CRESS DNA) viruses. A large number of diverse CRESS DNA viruses has been discovered in environmental samples, such as animal faecal material (Blinkova *et al.*, 2010; Kim *et al.*, 2012; Sikorski *et al.*, 2012; van den Brand *et al.*, 2012), water (Labonté & Suttle, 2013; López-Bueno *et al.*, 2009; Ng *et al.*, 2012; Rosario *et al.*, 2009a; Roux *et al.*, 2012), sewage (Cantalupo *et al.*, 2011; Ng *et al.*, 2012), air (Whon *et al.*, 2012), soil (Kim *et al.*, 2008) and ocean sediment (Yoshida *et al.*,

2013). Interestingly, insects have also been a valuable source for investigating ssDNA viral diversity, for example a study undertaken by Rosario *et al.* (2013) identified ssDNA viruses in the gut of dragonflies with the rationale that dragonflies prey on insects that vector geminiviruses. In this study a new species of mastrevirus (DfaMV) and an alpha satellite were isolated from dragonflies (*Erythrodiplax fusca* and *Erythrodiplax vesiculosa*) from Puerto Rico. Since this study many CRESS-DNA viruses have also been isolated from dragonflies (Rosario *et al.*, 2012a; Rosario *et al.*, 2011). Other studies have looked at insect vectors such as mosquitoes and whiteflies which feed on animals and plants, respectively and these have proven to be a useful way of sampling ssDNA viral diversity circulating in a region (Ng *et al.*, 2011a; Ng *et al.*, 2011b).

Amongst the CRESS-DNA viruses identified over the past five years are ssDNA viruses whose Reps are most similar to geminiviruses. Many of these are members of the proposed gemycircularvirus genus. Gemycircularviruses have been isolated from a variety of sources which include, the fungi *Sclerotinia sclerotiorum* (Yu *et al.*, 2010; Yu *et al.*, 2013), river sediment (Kraberger *et al.*, 2013b), animal faeces (Sikorski *et al.*, 2013), dragonflies (Rosario *et al.*, 2012a), mosquitoes (Ng *et al.*, 2011b) and plant material (Dayaram *et al.*, 2012; Du *et al.*, 2014). SsHADV-1, the first member of the gemycircularviruses infects the fungus *S. sclerotiorum* and is the only member whose host has been identified. Based on this and the fact that numerous Rep-like sequences have been identified in fungal genomes that are most closely related to gemycircularviral Reps, it is postulated that other members may also infect fungi. Other CRESS-DNA viral genomes have been isolated which are most closely related to geminivirus and the gemycircularviruses, these have been recovered from raw sewage material and have been named Baminivirus, Niminivirus and Nephavirus (Ng *et al.*, 2012). Several studies have collected NGS data on environmental samples reporting that a portion of the resulting DNA fragments share similarity to geminiviruses, these studies however have not recovered full viral genomes (Kim *et al.*, 2008; López-Bueno *et al.*, 2009; McDaniel *et al.*, 2013; Rosario *et al.*, 2009a; Soffer *et al.*, 2013; Whon *et al.*, 2012). The discovery of these geminivirus-like viruses as well as a rich diversity of other CRESS-DNA viruses in environmental samples highlights the fact that viral diversity of ssDNA viruses has been greatly underestimated.

1.7 Aims and rational of this study

The main objective of this thesis was to gain insight into the dynamics of mastreviruses by investigating the diversity, host range, geographic distribution, evolution and global movements of mastreviruses. In addition to this, use a viral metagenomic approach to identify mastreviruses or similar viruses that may be present in wild *Poaceae* sp. and treated sewage material in New Zealand.

The focus of mastrevirus research has predominantly been on two mastrevirus species, MSV and WDV, due to the economic impact these two viruses have on maize and cereal crops, respectively. Additionally some studies have investigated diversity and recombination patterns of wild grass-infecting mastreviruses such as PanSV (Varsani *et al.*, 2009a; Varsani *et al.*, 2008a), however, little is known about the diversity and dynamics of wild grass-infecting mastreviruses. Only one strain of MSV, MSV-A is known to infect maize. This maize adapted strain also infects grasses and has a wide geographical range within Africa. Recombination patterns have been documented among MSV strains and there is evidence that ancestors of MSV-A variants were the result of recombination events between MSV-B and MSV-G/F variants (Varsani *et al.*, 2008b). It is therefore important not only monitor viral populations in crops but also those in weeds that may act as viral reservoirs where mixed infections can occur and facilitating recombination. By understanding natural diversity we can be better equipped for dealing with emerging viral pathogens.

The majority of monocot-infecting mastreviruses identified outside of Africa have been found in Australia. Prior to studies undertaken as part of this thesis, four species originating from Australia have been identified and only a single isolate of each was available in the public databases. In Chapter Two monocot-infecting mastreviruses infecting wild *Poaceae* sp. in Australia is researched, to investigate for the first time mastrevirus dynamics in Australia.

The study carried out in Chapter Three builds on previous studies (Shepherd *et al.*, 2008b; Varsani *et al.*, 2009a; Varsani *et al.*, 2008a; Varsani *et al.*, 2009b; Varsani *et al.*, 2008b) by undertaking an extensive survey of mastreviruses infecting wild *Poaceae* sp. in Africa and

the SWIO and aims to expand on current knowledge of monocot-infecting mastrevirus diversity, host range and evolution through recombination.

Among dicot-infecting mastreviruses, CpCDV has been the most extensively studied due to the agricultural impact to pulses in the major growing regions of the Middle East, north-east Africa, Pakistan and India. Most of the studies on CpCDV prevalence and host range have used serological assays and only a handful of full genomes were available on GenBank prior to the work undertaken in this thesis. As a result little information regarding the diversity of dicot-infecting mastreviruses globally, including those species found in Australia which are distinct from CpCDV, was available. Studies have dated the origin of the mastrevirus MSV-A and highlighted the possible movements of this virus throughout Africa (Harkins *et al.*, 2009b; Monjane *et al.*, 2011b). In Chapter Four analysis of dicot-infecting mastrevirus samples spanning 27 year period from eight countries (Australia, Eritrea, India, Iran, Pakistan, Syria, Turkey and Yemen) was undertaken to gain a better understanding of the most likely geographic origin of the dicot-infecting mastreviruses and subsequent global dispersal.

Pakistan is a major pulse growing region where CpCDV has been documented. In Chapter Five an attempt is made to investigate CpCDV diversity within a major pulse growing region to further elucidate CpCDV strain dynamics on a more localised scale.

Building on the work undertaken in Chapter Five a further attempt is made to investigate CpCDV diversity and dynamics within a country in Chapter Six. Chapter Six is a comprehensive investigation of CpCDV within the major pulse growing regions of Sudan. In this study more than 140 CpCDV genomes isolated from pulse material are examined to allow some important insights into dicot-infecting mastrevirus dynamics, strain diversity on a regional scale and investigations into CpCDV evolution.

Geminiviruses are found in all growing regions of the world. In New Zealand however, there have been no reports of geminiviruses other than of two species which infect ornamental plants, *Abutilon mosaic virus* and *Honey suckle mosaic virus* (Lyttle & Guy, 2004). The

question on the presence of geminiviruses in New Zealand coupled with the recent discoveries of several gemini-like ssDNA viruses in a variety of sample types (Dayaram *et al.*, 2012; Labonté & Suttle, 2013; Ng *et al.*, 2012; Ng *et al.*, 2011b; Sikorski *et al.*, 2013; Yu *et al.*, 2010) prompted the study described in Chapter Seven. In this study a viral metagenomic approach was used to investigate the presence of potential mastreviruses or related ssDNA viruses in wild *Poaceae* sp. been shown to harbour a rich diversity of mastreviruses, particularly in Africa and Australia (Kraiberger *et al.*, 2012; Martin *et al.*, 2011c; Shepherd *et al.*, 2008b; Varsani *et al.*, 2008a; Varsani *et al.*, 2009b) and it is therefore highly plausible that there may be mastreviruses or similar viruses present in New Zealand grasses. New Zealand also cultivates many monocot crops such as maize, wheat and barley which are potential naive hosts for emerging viruses and therefore it may be prudent to gain insights to what potential viral pathogens are present in native/endemic *Poaceae* populations.

In Chapter Eight treated sewage material is used as a sample source to identify mastreviruses or novel gemini-like viruses. The rationale behind investigating sewage material is that humans consume a variety of plants which harbour plant viruses. The RNA virus *Pepper mild mottle virus* has been shown to be highly prevalent in wastewater across the USA and as a result this virus has been proposed as a possible indicator of human faecal contamination in water systems (Rosario *et al.*, 2009b). This demonstrates that plant viruses can be identified in sewage material and therefore it may be possible to use sewage material to investigate local ssDNA viral populations, particularly those potentially infecting plants.

	Motif I	Motif II	
FR687959 CpCDV-A-----	MPSANKNFRFQSKYVFLTYPKCSSQRDALLEFLWEKLT-PFLIYFIGVATELHQDGTTHYHALIQLDKRPHIRDP	SFF	[120]
JN989415 CpCV-A-----	MPSSSKRQNNFRLQTKYVFLTYPHCSSSTATSLRDFLWEKLS-RFAIFFIAVATELHQDGTPLHLHCLQLDKRGDIRDP	SFF	[120]
JN989420 CpCAV-----	MPQTKKPSSSFRLQTKYVFLTYPRCSSDAESLRDFLWEKLS-RFAIFFIAVATELHQDGTPLHLHCLQLDKRSNIRDP	SFF	[120]
GU256532 CpRLV-----	MPRLNKKTSNFRFQSKYVFLTYPHCNSNPEALRDYLWEKLT-RFIIFFIAVASEVHQDGSPLHLHCLQLTNKPNISDAS	SFF	[120]
JN989439 CpYV-----	MPSPSKKSPSFRQLQTKYVFLTYPHCSSSAEGLRDFLWDKLS-RFAIFFIAIATELLPA---HLHCLQLDKRSNIRDP	SFF	[120]
M81103 TYDV-A-----	MPSAPQKTKSFRQLQTKYVFLTYPRCSSSAENLRDFLWDKLS-RFAIFFIAIATELHQDGTPLHLHCLQLDKRSNIRDP	SFF	[120]
Y00514 MSV-A-----	MASSSSNRQFSHRNANTFLTYPKCPENPEIACQMIWELVV-RWIPKYILCAREAHKDGSPLHLHALLQTEKPVRI SDS	RFF	[120]
JQ624879 MSRV---MSAFGNHFVHMPSVQAGVFNPMPGSDYPSEEDLHQNTPVGPDPPRRNFQKSANAF	FLTYPRCLLTPEAGQHLWEVAR-HWTPSYVLASSESHQDGTPLHLHVM---RPMSTRDP	SFF	[120]
KJ437671 ACSV-----	MNTEHGGPSGFRFQSRNIFLTYPRCNLAPELIGSFLLSLLS-PYHVMFITVTSELHKDGTPLHIALAQTDKRVHTYSP	GFF	[120]
L39638 PanSV-A-----	MSTSLSITSDGRHSVRSFRHRNANTFLTYSKCPLPEFIFEHLFRLTK-DFEPAYILVRETHQDGTWHCHALLQCIKPVTTTRDE	RYF	[120]
AF072672 SSRV-A-----	MPSQEDSTVASRPFKHRNANTFLTYSRCRLDPEAVGLILWQLIS-HWSPAYILVSREAHADGEWHLHALVQSVRPVQTNTQ	GFF	[120]
M82918 SSV-A-----	MSTVGSTVSSTPSRRFKHRNVNTFLTYSRCPLPEAVGLHIWSLIA-HWTPVYVLSVRETHEDGGYHIHVLAQSAKPVYTTDS	GFF	[120]
EU244915 ESV-----	MSSIASTVPSAPTTRRFKHRNVNTFLTYPHCTLEPEVGLVLSLLE-SWTPAYIIVSREAHQDGSWHLHALAQSVKPVYTHDER	RFF	[120]
EU445697 USV-----	MATVGSSSNSVASRSFKHRNANTYLTYPKCPLPEAIGLTLWSLIA-PWEPAYIIVCREAHQDGTWHCHALAQSVKPVTTTRNS	RFF	[120]
FJ665632 ECSV-----	MASSSHFRFIQGRAFLTYSQCPREP KDVGEFLTSHSTLASHVYVVRVQKEHQDGNHLHAIVCTSERDIRDP	RIFF	[120]
GQ273988 SacSV-----	MAYANSTSTESNSSRSFRHRNANTFLTYSKCCLDPEILGLSLWSKLA-PWTPAYILVAREAHQDGTWHCHALAQSVRPVTTSDP	RFF	[120]
JF508490 EMSV-----	MSQDQDTLGSSSDGSRFRISSKQLFLTYPRCDLSPKDLGLELLQLLI-QNKPKYIHVQELHKDGFPLHALVQLEKKLFTRRQ	TFF	[120]
AF239159 SSEV-----	MTTVGSAESGSAIRSFKHRNVNTFLTYPKCHLEPEAVGLHLWSLIG-HWNPAYIIVSREAHADGSWHIHALAQSVKPVQTNTNP	RFF	[120]
KJ210622 SWSV-----	MSTESSNEGIAAQRSLTGLEEFFTTTWPDPFRAGSTSNPRVFQFKAQNIFLTYPRCDISVDVAARNLLTLCH-RFQPLYILCSQEHHADGSN	HLHILLQTDKTM YTRNPHYF	[120]
JQ948091 DCSMV-A-----	MAPPVSDTESANSANCRLAERAPGAAEASFVRAKNIFLTYSKCLLDPEALRDITHKLR-KFEPTYVYVARELHQDGTFLHLCFVQCKKHVRTTR	RFF	[120]
JF905486 PSMV-A-----	MSSLVSETSNSEVGSQMESPGRGGQSIDAPSSSCFKVRARNLFLTYSKCNLTAVFLLEYISSLLK-KYCPTYIYVAQEAHKDGS	HLHCI IQCSKYVRTTSKFF	[120]
M20021 CSMV-----	MSSLPVSESEGEGSTSVQVPSRGGQVTPGEKAFSLRTKHVFLTYPRCPISP EEAQQKIADRLK-NKKCNYYISREFHADGEP	HLHAFVQLEANFRTTSPKYF	[120]
JQ948051 SSMV-1-----	MSGPSRPPSPFAISSSDEESVDGFHFRGKNIFLTYSRCEIDPALITDALWDKFS-SHKPLYILSVRELHQDSGFVHVCLVQLTDQYRSRDS	SFA	[120]
JQ948052 SSMV-2-----	MSSQSNSTEASPANFRFRARSFLTYPKCTLEPRDVVEHLYSKFR-KYGP KYCLVTR EHS DGDYHLHCLFQLDKAFSTNDS	SFF	[120]
HQ113104 BCSMV-----	MASFVSETSDARGQTGAPRSPSGEVGAPGAVAACFEVRSRNIFLTYSKCHLDPVFMQEHLSSLR-RFEPTYVYVAREEHQDGSY	HLHCLVQCKKYVRTKSAKFF	[120]
HM122238 DDSMV-----	MSSQLVSDSVMFDP RSYGEYPSSESAASLPSFNVRSQHVFLTYPRCPIPPKDAGSFLKKLCK-RYNIQYMYIAQELHQDGEPLHAF	FLQFDKVFRTTSAKYF	[120]
JQ948087 PDSMV-----	MASHVSETEGARGQVGAPPLQGEVVGAPGAVEACFEVRSRNVFLTYSRCHLEPSFMLERLSRLK-KWDPTYSYVAREEHKDGSY	HLHCLVQCRKYIRTKSAKFF	[120]
AJ783960 WDV-A-----	MASSAPRFRVYSKYLFLTYPQCILEPQYALDSLRTLLA-KYEPLYIAAVRELHEDGSPHLHVLVQNKLRSITNP	NAL	[120]
JQ361910 WDIV-----	MSQ TSAENNSANPKASSSTFRYRSNNCFLTFPHCNSCPYGMVQHFWDLIS-TWSPYIYAVASVELHQDGTPLHLALLQTRKQISTNDP	HFF	[120]
AM296025 ODV-----	MATVASSSTRFRVYSKYLFLTYPQCILEPQYALDSLRLQ-KYKPLYICSVRERHEDNSPLHLHVLVQCEKRASITNP	NAL	[120]
M23022 DSV-----	MAANRSFRHRNANTFLTYSKCDHSPQLIADHLWDLK-SWNPIYILVASEHHADGSLSHSHALVQTEKQVNTNTQ	RFF	[120]
E02258 MiSV-----	MRAPASSAASNRPGPSNHPTPRWNSKQFLTYPHCNLTPELMKELFSRLT-EKIPGYIKVSQEFHKDGDPLHLHVLQLNTKLCTRN	KFF	[120]
JX458741 DfasMV-----	MSSSSAQSGPSHFIRAQNIFLTYPRCDLDPKDAGEIIQSKMQ-SHEPKYILFSRELHSDGEYHLHGLQLLSRQFSSNNP	RIFF	[120]

Additional figure 1.1 (continued below)

	GRS Domain	Motif III	
FR687959 CpCDV-A	D-----FEG--NHPNIQPARNSKQVLDYISKD	---G-DIKTRGDF--R-----DHKISPSK-----SDARWRTIIQTATTKEEYLDMIK	[240]
JN989415 CpCV-A	D-----FEG--NHPNIQPAKNSEQVLDYISKD	---G-NVITRGDF--R-----KHKVSPTK-----HDKRWRTIIQTATTKEEYLGMIK	[240]
JN989420 CpCAV	D-----FQG--NHPNIQPAKNSEQVLDYISKD	---G-SVITRGDF--R-----KHKVSPTK-----SDDRWRTIIQTATTKEEYLEMIK	[240]
GU256532 CpRLV	D-----FEG--NHPNIQPARNSKQVLDYISKD	---G-NIITKGEF--K-----KHRVSPTK-----HDERWRTIIINTATSKEHYLGMIK	[240]
JN989439 CpYV	D-----FEG--NHPNIQPAKNSEQVLDYISKD	---E-NVITKGEF--R-----KHKVSPTK-----TDERWRNIINTATSKEEYLGMIK	[240]
M81103 TYDV-A	D-----LEG--NHPNIQPAKNSEQVLEYSKID	---G-NVITKGEF--K-----KHRVSPTK-----SDERWRTIIQTATSKEEYLDMIK	[240]
Y00514 MSV-A	D-----ING--FHPNIQSAKSVNRVRDYILKE	---PLAVFERGTFIPR-----KSPFLGKSDSEVKEKKPSKDEIMRDIISHATSKAEYLSMIQ	[240]
JQ624879 MSRV	D-----IQG--YHPNIQASRS PNKTREYILKS	---PITVYSRGTFIPR-----AGTSGAGY---GSTPVPKRNEIMRGI IETTNNKAEYLSEVQ	[240]
KJ437671 ACSV	D-----VQG--FHPNIQSARS PQTVLSYILKS	---PTGTFNYGSLRPRGTRADAACGIGEDTAGGGASSSSPPQADQRRGSRRELANDPGRDRKDVLMTSILAGSSSKQEFNLGVK	[240]
L39638 PanSV-A	D-----IDR--YHPNIQSAKSTDKVREYILKD	---PKDKWEKGTIYIPR-----KKSFPVPPG--KENSEKKPSKDEVKMEIMTHATSRAEYLSLVQ	[240]
AF072672 SSRV	D-----IES--FHPNIQSAKSANKVREYILKN	---PIAKWEKGTIYIPR-----KQCFVSSS--SESNSKPSKDDIVRDIIEHSTSKEEYLSMLQ	[240]
M82918 SSV-A	D-----IDG--FHPNIQSAKSANKVRA YAMKN	---PVTYWERTGTFIPR-----KTSFLGDS--TEPNSKKQSKDDIVRDIIEHSTNKQEYLSMIQ	[240]
EU244915 ESV	D-----IED--YHPNIQAAKSANKVRDYVLKN	---PLKVWERTGTFIPR-----KKTFLGST--SEGNTTKQSKDDIVRDIIEHSTSKQEYLSMIQ	[240]
EU445697 USV	D-----IED--HHPNIQSAKSVDKVRA YILKD	---PIALWERTGTFIPR-----KKSFPVHQ--GDEHTPKPTKDDIVRDIIEHSTSKQEYLSRLQ	[240]
FJ665632 ECSV	D-----FGE--FHPNIETCRSVSKSLKYIQKE	---AGSFYEHGT--VPC-----DKRLTGRK-----RKAEQDEWWHQAVNSG--SIEEALQLVK	[240]
GQ273988 SacSV	D-----VNE--YHPNIQSAKSVDRVREYILKD	---PLCQWEKGTIVPR-----KKPFVPQI--GESSNTRASKDDIVRDI IQHSTNKHEYLSMLQ	[240]
JF508490 EMSV	D-----QFLHGTFKHPNIQPARDA SKVLGYITKQ	---NGEYIFGKPTLP-----KKKKTAEQ-----GRDQRMRAIIESSTSKQEYLSMVR	[240]
AF239159 SSEV	D-----IED--FHPNIQSAKSADRVKE YVLKN	---PIKQWEKGTIYIPR-----KKSFATTS--SEDRQPKPTKDDIVRDIIEHSTSKQEYLSMIQ	[240]
JQ948091 DCSMV-A	D-----LEE--YHPNVQNARMPHKVLA YCKKS	---PVSYAEEGAYTES-----DVRKKRID-----ASTTKDAKMADIIRSSKSKEEYLSMVR	[240]
JF905486 PSMV-A	D-----IKE--FHPNVQNPRMPKKALS YCKKS	---PISEAEYGVFQEI-----KRPRKKKA----DAPSTKDAKMAEIIKSSTNKEDYLSMVR	[240]
KJ210622 SSNV	D-----ICG--HHPNIQPAKSPDNVRA YILKD	---PITSFEEGSGFQPR-----GSRNSTSA--IPRSGNSGTDKSLMRDIINTSTSKDDYLTRVR	[240]
M20021 CSMV	D-----LDE--FHPNIQAARQPASTLK YCMKH	---PESSWEFGKFLKP-----KVNRSPTQ-----SASRDKTMKQIMANATSRDEYLSMVR	[240]
JQ948051 SSMV-1	D-----LGG--NHPNIQTVRSATKVKE YILKE	---PVSQSARGKFVAP-----GGRPPKHTDRRRSDSAVKDERMRYILRTATTRDDYLGMR	[240]
JQ948052 SSMV-2	N-----ILD--YHPNIQTAKSPTNVRDYCLKN	---PVSKAERTGTFIPL-----KGRTPKNT-----ESKAKDSVMRSIIINTSTDRASYLSMVR	[240]
HQ113104 BCSMV	D-----VEE--FHPNVQNARMPHKVLA YIKKN	---PLCFVETGVFQAS-----TKQKKKKV----DAPSTKDAKMAEIIKSSTCKEDYLSMVR	[240]
HM122238 DDSMV	D-----FFE--FHPNIQAARNPEKTLE YCQKN	---PADFYEDGVFVKP-----KASRKRKL-----ASFTRDKMKQIMANATSRDEYLSMIR	[240]
JQ948087 PDSMV	D-----VEE--FHPNVQNARVPHKVLA YIRKG	---PVCFVEHGAFKDE-----AKEKKKKA----DAPSTKDAKMASIISQSTSREYDLGMVK	[240]
AJ783960 WDV-A	NLRMNTSPFSI--FHPNIQAAKDCNQVRDYITKE	VDSVDVNTAEWGTFIIV-----TTPGR-----KDRDADMKQIIESSTSREEFLSMVC	[240]
JQ361910 WDIV	D-----FDG--HHPHIQAAKNPTLCRDYILKG	---PITFSEKGAFIPR-----GRNAGTSP-----RHSNKRSRDDIMKDIIENSTNKS DYLSKVR	[240]
AM296025 ODV	NLRMDLSPFTT--YHPNIQPANN CN DVREYITKE	VDSHEHTAEWGTFIHH-----TTSRGPDK-----DEAMKQIIESATSKEEFISMVR	[240]
M23022 DSV	D-----ILE--FHPNIQSAKSVNKRVTYILKN	---PVEKFERGTFVPR-----KSPFLGES--SSSEKHNKDDVMRDIIDHATSKEEYLSMVQ	[240]
E02258 MiSV	D-----VQG--FHPNIQPVDAEKVFGYISKT	---NGDSDMGELQLR-----NKKPEK-----PTRDQRMAMIIASSTNRNEYLSMVR	[240]
JX458741 DfasMV	D-----IGA--HHPNIQSAISPKSVRDYILKN	---PITQFCIGTYVPA-----KKGRKLGS----RFEENIRNNIMRSIIISTATSKESYLSMVR	[240]

Additional figure 1.1 (continued below)

	BBR interaction domain	Walker-A	Walker-B	
FR687959 CpCDV-A	EEFPHEWATKLQWLEYSANKLFPPQPEPYVSPFT--ESD LRCH EDLAAWRDKHLYHDDGRSG-IRHPSLYIC GPTRTGKTT WARS LG--RHNYWNGTIDF--TTYDDHATYNI IDD IPFK [360]			
JN989415 CpCV-A	DQFPHEWATKLQWLEYSANKLFDPDIEPPYENPFS--PID LQCH EIQEWLNRDLYVEPEQLQ-HRRNSLYIC GPTRTGKTS WARS LG--RHNYFNGGVDF--TTYDINATYNI IDD IPFK [360]			
JN989420 CpCAV	TEFPHEWATKLQWLEYSANKLFDPDIEPPYSSPFP--NEF LQCD EITEWLNRDLYQEPEQLQ-HRRNSLYIC GPTRTGKTT WARS LG--RHNYFNGGIDF--TTYDPNATFNI IDD IPFK [360]			
GU256532 CpRLV	DQFPHEWASKLQWFEYSANKLFDPDVEPPYQSPFP--EAS LQCH EIQDWLNRDLYLEPEQLR-HRRKSLYIC GPTRTGKTS WARS LG--RHNYFNGGVDF--TTYDEQAAYNI IDD IPFK [360]			
JN989439 CpYV	EEFPHEWATKLQWLEYSANKLFPEIEPPYQSPFT--SCS LQCH EKITDWLNRDLYLEPEQLR-HRRNSLYIC GPTRTGKTS WARS LG--LHNYFNGGVDF--TTYNPLATYNI IDD IPFK [360]			
M81103 TYDV-A	DEFPHEWATKLQWLEYSANKLFPPQPEIYQATFT--EED LQCH EDLQLWRDQHLHYHEPRRAG-TRIPSLYIC GPSRTGKTT WARS LG--RHNYWNGTIDF--TVYDDHATYNI IDD IPFK [360]			
Y00514 MSV-A	KELPFDWSTKLQYFEYSANKLFPEIQEEFTNPHPPSSPD LLCNE SINDWLQPNIFQSSDER--SRKQSLYIV GPTRTGKTS WARS LG--VHNYWQNNVDW--SSYNEDAIYNI IDD IPFK [360]			
JQ624879 MSRV	KAFPFEWATKLQQFEYSAERLFPTLPSPFVPPHPSPSEP LHCYE TIRSWKDENIFQGDAASTRSRPSLYIV GPTRTGKTT WARS IDPVNHNYWQNGVDF--LKyrksAKYNI IDD IPFK [360]			
KJ437671 ACSV	KAFPYDFCARLQNW EYAANKLF-DTPAVYQPPFP--DSY FHCH ENIHWDVRDNIYEITPE---ARPLSLYIC GPTRTGKTS WARS LG--RHNYWQNNVDF--TSYDVEAKYNI IDD IPFK [360]			
L39638 PanSV-A	TSLPYDWATKLSYFEYSASRLFPDIAEPYSNPHPATDPD LLCNE TLQDWLEPNYQIIPG---ARKRSLYIV GPTRTGKTS WARS LG--RHNYWQNNVDW--SSYDEEAAYNI IDD IPFK [360]			
AF072672 SSRV-A	KALPYDWATKLQYFEYSASKLFDPDVEEYTSPPHTTTPL LRDPT ITIDNWVQPNLFQNNTG---TRKLSLYIL GPTRTGKTS WARS LG--RHNYWQNNVDW--SCYDEDAVYNI IDD IPFK [360]			
M82918 SSV-A	KALPYEWATKLQYFEYSANKLFDPDIEIYTSPPFPQSTPA LLDPT AINTWLENNLYQNSNS---NRKLSLYIL GPTRTGKTS WARS LG--RHNYWQNNVDW--SSYDEDAEYNI IDD IPFK [360]			
EU244915 ESV	KALPYEWATKLQYFEYSANRLFPEIQETYTNPHPTAPQ LQDGE TIQSWVYTNIYQNIIPG---TRKQSLYIL GPTRTGKTS WARS LG--RHNYWQNNVDW--TSYDEDAVFNI IDD IPFK [360]			
EU445697 USV	NELPYEWATKLQYFEYSANKLFDPDIEPEYIHPHPQTEPE LHCKE TIDDWLKPNIQQLPS---DRKQSLYIV GPTRTGKTS WARS LG--RHNYWQNNVDW--TSYDEEAMYNI IDD IPFK [360]			
FJ665632 ECSV	DNEPRTFWLQHHNLVTNARRIWSEVRAEFVPKYS--ESS FSVPR VLSDWVANNLRADPLP---DRPLSLIIE GD SRTGKTAWARS LG--RHNYLSGHLDLNGAVFDNEASYN IDD VNPK [360]			
GQ273988 SacSV	KELPYEWATKLQYFEYSANKLFPEIAEPYTNPHPTQPD LHCYE RIEEWLNFNVYQQPQEA--GRARSLYIV GPTRTGKTS WARS LG--RHNYWQNNVDW--SSYDEEAVLNI IDD IPFK [360]			
JF508490 EMSV	KEFPFDWATRLMQFEYSASSLFPPEPPVEYTSPPF--VDQ LLCPE DITEIINSEWFQHGAPG--GRPRSIYIC GPTRTGKTT WARS LG--RHNYNSVLDL--THYDPQAEYNI IDD VNPK [360]			
AF239159 SSEV	KALPYEWATKLQYFEYSASRLFPETAEVYTNPHPTPED LINFET IEDWLNPNYQNIIPG---HRKQSLYIL GPTRTGKTS WARS LG--RHNYWQNNVDW--SSYDEEAAYNI IDD IPFK [360]			
KJ210622 SSVN	NTFPFDWATRLQQFEYSASKLFPEPVREYVNPFPSEP LFCRE IIDRWIDITDAFDAA----QRRRSLYIV GPTRTGKTS WARS LG--RHNYWQHMDVDF--TAYDTHAKYNI IDD VNPK [360]			
JQ948091 DCSMV-A	KTFPFDWATRLQNF EYSAERLFPSTPPPYVSPFN--MPS QEEHP VLGAWLRAELYTGRNPA---ERRKSLYIC GPTSTGKTT WARS LG--KHNYWQHSVDF--LNIIPDAEYNI IDD IPFK [360]			
JF905486 PSMV-A	KSFPPDWATRLQQFQFSAESLFPSTPPPYVDPFG--MPS QDTHP VIGAWLRDELYTDRSPT---ERRRSLYIC GPTRTGKTS WARS LG--SHNYWQHSVDF--LHVIQNARYNI IDD IPFK [360]			
M20021 CSMV	KSFPPFEWAVRLQQFQYSANALFPDPPTYSAPYA--SRD MSDHP VIGEWLQQELYTVWSPG---VRRRSLYIC GPTRTGKTS WARS LG--THHYWQHSVNFL--EENWCQAQFNI IDD IPFK [360]			
JQ948051 SSMV-1	KSFPPFEWATRLAQFEYSASKLFDPDITPQYQSQYQ--TTD LTCH ENLLDWYQENLCYIDGA---GRRKSLYIC GPTRTGKTS WARS LG--RHNYYNMQVDW--ATYDQEAQYNI IDD IPFK [360]			
JQ948052 SSMV-2	KAFPFDWATKLQQFEYSASKLFDPDVIPEYTSPPF--TEN LMCNE RITDWLNTLYSADHPR---TRKSGLYIC GPNTGKTS WARS LG--KHNYWQMNLDL--ANYNNEAQYNI IDD IPFK [360]			
HQ113104 BCSMV	NTFPFDWATRLQQFQYSAESLFPVPTPYMDPFG--MPA QDEHP VIGAWLQALFSDRRPD---ERRRSLYIC GPTRTGKTS WARS LG--AHNYWQHSVDF--LNLVANATYNI IDD IPFK [360]			
HM122238 DDSMV	KAFPFDWAI RLQQFEYSAKALFPEAPIQYQPQFV--SND MSDHP VIGEWLDTEFFTERGPH---HRRRSLYIC GPTRTGKTS WARS LG--THHYWQHSVDL--TEWNKNAIYNI IDD IPFK [360]			
JQ948087 PDSMV	KEFPFDWATRLQQFEYSAQALFCLPPPVDVDF--MPS QAEHQ VLGAWLREELYSDRSPA---ERRRSLYIC GPTRTGKTT WARS LG--CHNYWQHSVDF--LHVIPTARYNI IDD IPFK [360]			
AJ783960 WDV-A	HRFPFEWSIRLKDFEY TARHLFPDPVNTYTPEFP--IES LMCH ETIESWKNEHLYSSP-----GRHKS IYIC GPTRTGKTS WARS LG--IHNYNSLVDF--TTYDVNAKYNI IDD IPFK [360]			
JQ361910 WDIV	RNFYDWATKLYNFEYSASKLFPEQQPEYSNPHGQSVPD LYCYE TIQDWIDSNLFQDPSAG---TRPKSLYIV GPTRTGKTS WARS LG--RHNYWQNNVDF--TVYDPEAAAYNI IDD IPFK [360]			
AM296025 ODV	SRFPFEWSINLQRFQYTANYLFDPDIPQYTPEFP--TES LICH ETIQNWANTELFTV-----RRHRSLYIC GPTRTGKTS WARS LG--IHNYNSQVDF--TNYNADALYNI IDD IPK [360]			
M23022 DSV	KALPYDWATKLSYFEYSADRLFPVEAAPFINPHPPSEP LLCQE TIIDWLQNDLFQVVTDG---VRKRSLYIL GPTRTGKTS WARGLG--RHNYWQNNVDW--ASYDEEAQFNI IDD IPFK [360]			
E02258 MiSV	KEFPFDWAI RLQQFEYSAALFTEPPPVYQSPFP--NEQ IVCP PELVDIIDQEWNPNGP---RRPRSIYIC GPSRTGKTT WARNIG--RHNYNSTVDF--THYDKDAIYNI IDD VNPK [360]			
JX458741 DfasMV	KSFPPFEWATKLSQFEYSASKLFPEVTPEYKSPFP--TES LCNE NIQDWVDNTLYQPNRT---SRGLSLYIC GPTRTGKTS WARS LG--VHNYWQNNIDF--SVYNDNATYNI IDD IPFK [360]			

Additional figure 1.1 (continued below)

Motif C

```

FR687959 CpCDV-A FVPLWKQLIGCQSDFTVNPYGGKKKI-KGGIPSIILCNPDDEDW---VPSMSSQKEYFEDNCITHYMSGDNF----FA-----RESSSH----- [ 468]
JN989415 CpCV-A FCPNWKQLVGSQKDFTVNPYGGKKRI-KGGIPCIILVNDDDDW---LLDMSSSQKEYFESNYKIHYMDSEETF----IA-----PESSSH----- [ 468]
JN989420 CpCAV FCPNWKQLVGSQKDFTVNPYGGKKRV-KGGIPTIILVNDDDDW---IKDMSSSQNEYFESNCLIHYMTEGETF----IA--ARRQVTQRASPPSENI----- [ 468]
GU256532 CpRLV FCPNWKQLVGSQKDFTVNPYGGKKRI-KGGVPCIILVNDDDDW---MKDMSSHQKEYFQHNCMIHYMDEGETF----IA-----PVSSSH----- [ 468]
JN989439 CpVY FCPNWKQLIGSQKDFTVNPYGGKKRI-KGGIPCIILVNDDDDW---MNDMSPAQNDFYQANCCIHMEEGETF----IA-----LQSSSH----- [ 468]
M81103 TYDV-A FVPLWKQLIGCQSDFTVNPYGGKKKI-KGGVPCIILCNDDDEDW---LKNMSPAQIEYFEANCITHFMYAAETF----FA-----PESSSH----- [ 468]
Y00514 MSV-A FCPCWKQLVGCQRDFIVNPYGGKKKVQKSKPTIILANSDEDEDW---MKEMTPGQLEYFEANCIYIMSPGEKW---YS-----PPVLPPTTEEV----- [ 468]
JQ624879 MSRV FCPCWKQLVGGQKDYTVNPYARRMEV-PGGIPSIILVNYDEDW---LKVMTPAQLEYFYDNCVVYQMEHKEF---YT-----PS----- [ 468]
KJ437671 ACSV YCPCWKALIGGQKDFTVNPYGGKKLI-KGGIPSIIVLNDDDEDW---MRAMTASQRSYFERNVCVVYLYEGDSF---IK-----DDVSSTSEECI----- [ 468]
L39638 PanSV-A FCPCWKQLVGCQKDYIVNPYGGKKRVASKSIPTIILANEDDEDW---LKDMTPAQYDYFYANCEIYVVMQAGEKW---FT-----PA----- [ 468]
AF072672 SSRV-A FCPCWKQLIGCQENYVNPYGGKKRVAKKSISTIILANEDDEDW---MKVMSPQLDYFHQNCVVYIMEEGERF---FG-----GPAVSATAHPPIGV--- [ 468]
M82918 SSV-A YCPCWKQLIGCQKDYIVNPYGGKKRVASKSIPTIILANEDDEDW---LRDMTPAQQDYFNANCETYMLEPGERF---FS-----LPAVSATAHPSSEV--- [ 468]
EU244915 ESV FCPCWKQLIGCQKDYIVNPYGGKKRVAKKSIPITIVLANVDEDW---LKDMTPAQQDYFNANCTVYILEPGERF---FG-----GPAVSATAHPSIEV--- [ 468]
EU445697 USV YCPCWKQLIGCQKEYIVNPYGGKKRVASRSIPTIILANEDDEDW---LKDMTPAQREYFEANCVIYIMTPGEKW---FS-----PV----- [ 468]
FJ665632 ECSV YLKHWEFIGAQKDWQSNLKYGKPVLV-KGGKPAIVLCNSDQSYKSFLDCEENHQLRSWTSKNALFVDIQDALFGGVSLTMREQTREDDPESPMWASDSDPGDQAV-- [ 468]
GQ273988 SacSV YCPCWKQLVGCQKNYVNPYGGKKKVAKRSIPAVILANEDDEDW---LRDMTPAQRDYMEANCEVYIMSSGEKW---FT-----PA----- [ 468]
JF508490 EMSV YCPQWKALVGAQRDYIVNPYGGKKKI-KGGIPSIILTNDDDEDW---LGEMKPAQAEYLHANSHVHYMYEGTKF---YK-----AEAAGQDV----- [ 468]
AF239159 SSEV FCPCWKQLVGCQKEYVNPYGGKKKVASKSIPSIILANEDDEDW---LTVMTPGQREYFEANAVIYIMTAGEKF---YK-----PA----- [ 468]
KJ210622 SSV FCPNWKQLVGCQRDFIVNPYAKRKEI-PGGIPCIILQNPDDEDW---LPVLSPSQMDYFVNNDVYVMKPGERF---FG-----GDPVPEAQEDVPDGTGSS [ 468]
JQ948091 DCSMV-A FVPCWKGLVGAQRDITVNPYGGKKRL-SNGVPCIILANEDDEDW---LQMQPGQADWFNANCEVHYMYQGETF---FK-----SLGAATA----- [ 468]
JF905486 PSMV-A FVPCWKGLVGSQKIDITVNPYGGKKRL-SNGIPCIILVNEDDEDW---LQMQPQSQADWFNANAVVHYMYSGESF---FE-----AL----- [ 468]
M20021 CSMV FVPCWKGLVGSQYDLTVNPYGGKKRI-PNGIPCIILVNEDDEDW---LQSMSTQQVDWFHGNNAVYHLLPGETF---IP-----SE----- [ 468]
JQ948051 SSMV-1 FCPHWKALIGCQKDFTVNPYGGKKLI-KGGIPTIILVNEDDEDW---LADMTPGQVSIFYEANVQIHYMTSEESF---IP-----DPALRQRLSLNYYKVCFFLM [ 468]
JQ948052 SSMV-2 FCPYWKALVGSQHEYTVNPYGGKKLI-KGGIPSIILVNEDDDW---MRAMNDGQRSYFEGNMSIYYMSEGESF---IR-----NEAL----- [ 468]
HQ113104 BCSMV FVPCWKGLVGCQFDITVNPYGGKKRL-KNGVPSIILVNEDDEDW---LQMQPQSQVGWGFETNCIIHYMYAGESF---FE-----A----- [ 468]
HM122238 DDSMV FVPCWKGLVGSQFDITVNPYGGKKTI-PNGIPSIILANEDDEDW---LQTMSPQQADWFHGNVCVVYLLQAGESF---IP-----PSSDVEA----- [ 468]
JQ948087 PDSMV FVPCWKGLVGAQREITVNPYGGKKRL-PNGIPCIILVNEDDEDW---PQQMQPSQAASFEDNCVVYFMNQGFRF---FE-----TTA----- [ 468]
AJ783960 WDV-A FTPNWKCFVGAQRDFTVNPYGGKKMI-RGGIPCIILVNPDEDW---LKDMTPAQSDYMSNAVHYMYEGESF---FA-----YGENVTASQ----- [ 468]
JQ361910 WDIV FCPCWKQLVAAQRDFTVNPYGGKKLI-KGGIPSIILVNSDEDW---LKTMTPEQQEYFEANSIYIMMEPTTEKF---FG-----GAEIV----- [ 468]
AM296025 ODV YVPNWKCFLGAQKDFTVNPYGGKKTI-RGGIPCIILVNPDDEDW---LKDMTPLQSDYLYANAEIHYMEDGETFINHSFT-----FGEGATASQ----- [ 468]
M23022 DSV FCPCWKQLIGCQKEYVNPYGGKKRVASKSIPSIILTNPDDEDW---MKDMTPAQLSYFEANTVIYKMTGEGERF---FS-----YAEGPATASLASLDDAPA-- [ 468]
EO2258 MiSV FLPQWKALVGAQRDYIVNPYGGKKKI-PGGIPSIILTNDDDEDW---IKDMKPAQVEYLYANAHVHYMYEGQKF---YV-----LPAEE----- [ 468]
JX458741 DfasMV FCPCWKALAGSQSDFTVNPYGGKKRI-KGGIPCIILVNEDDEDW---LTCMSSSQKTYFESNVVIYYMYAGEKF---FN-----FVEE----- [ 468]

```

Additional figure 1.1: Replication-associated protein annotation highlighting the functional motifs in representatives from each mastrevirus strain

1.8 References

- Akhtar, K. P., Ahmad, M., Shah, T. M. & Atta, B. M. (2011).** Transmission of chickpea chlorotic dwarf virus in chickpea by the leafhopper *Orosius albicinctus* (Distant) in Pakistan. *Plant Protection Science* **47**, 1-4.
- Akhtar, S., Khan, A. J. & Briddon, R. W. (2013).** A Distinct Strain of Chickpea chlorotic dwarf virus Infecting Pepper in Oman. *Plant Disease* **98**, 286-286.
- Alberter, B., Ali Rezaian, M. & Jeske, H. (2005).** Replicative intermediates of *Tomato leaf curl virus* and its satellite DNAs. *Virology* **331**, 441-448.
- Ali, M. A., Kumari, S. G., Makkouk, K. H. & Hassan, M. M. (2004).** Chickpea chlorotic dwarf virus, CpCDV naturally infects Phaseolus bean and other wild species in the Gezira region of Sudan. *Arab Journal of Plant Protection* **22**, 96.
- Amin, I., Hussain, K., Akbergenov, R., Yadav, J. S., Qazi, J., Mansoor, S., Hohn, T., Fauquet, C. M. & Briddon, R. W. (2011).** Suppressors of RNA silencing encoded by the components of the cotton leaf curl begomovirus-betasatellite complex. *Molecular Plant-Microbe Interactions* **24**, 973-983.
- Andersen, M. T., Richardson, K. A., Harbison, S.-A. & Morris, B. A. M. (1988).** Nucleotide sequence of the geminivirus chloris striate mosaic virus. *Virology* **164**, 443-449.
- Argüello-Astorga, G. R., Guevara-González, R. G., Herrera-Estrella, L. R. & Rivera-Bustamante, R. F. (1994).** Geminivirus Replication Origins Have a Group-Specific Organization of Iterative Elements: A Model for Replication. *Virology* **203**, 90-100.
- Argüello-Astorga, G. R. & Ruiz-Medrano, R. (2001).** An iteron-related domain is associated to Motif 1 in the replication proteins of geminiviruses: Identification of potential interacting amino acid-base pairs by a comparative approach. *Arch Virol* **146**, 1465-1485.
- Ashby, M. K., Warry, A., Bejarano, E. R., Khashoggi, A., Burrell, M. & Lichtenstein, C. P. (1997).** Analysis of multiple copies of geminiviral DNA in the genome of four closely related *Nicotiana* species suggest a unique integration event. *Plant molecular biology* **35**, 313-321.
- Atanasova, N. S., Roine, E., Oren, A., Bamford, D. H. & Oksanen, H. M. (2012).** Global network of specific virus–host interactions in hypersaline environments. *Environmental Microbiology* **14**, 426-440.
- Baliji, S., Black, M. C., French, R., Stenger, D. C. & Sunter, G. (2004).** Spinach curly top virus: a newly described curtovirus species from southwest Texas with incongruent gene phylogenies. *Phytopathology* **94**, 772-779.
- Bejarano, E. R., Khashoggi, A., Witty, M. & Lichtenstein, C. (1996).** Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. *Proceedings of the National Academy of Sciences* **93**, 759-764.
- Bergemann, A. D., Whitley, J. C. & Finch, L. R. (1989).** Homology of mycoplasma plasmid pADB201 and staphylococcal plasmid pE194. *Journal of bacteriology* **171**, 593-595.
- Bernardo, P., Golden, M., Akram, M., Naimuddin, Nadarajan, N., Fernandez, E., Granier, M., Rebelo, A. G., Peterschmitt, M., Martin, D. P. & Roumagnac, P. (2013).** Identification and characterisation of a highly divergent geminivirus: Evolutionary and taxonomic implications. *Virus Research* **177**, 35-45.

- Bigarré, L., Salah, M., Granier, M., Frutos, R., Thouvenel, J. C. & Peterschmitt, M. (1999).** Nucleotide sequence evidence for three distinct sugarcane streak mastreviruses. *Arch Virol* **144**, 2331-2344.
- Bigirwa, G., Gibson, R., Page, W., Hakiza, J., Kyetere, D., Kalule, T. & Baguma, S. (1995).** A new maize disorder in Uganda caused by *Cicadulina niger*. In), *Maize research for stress environments Proceedings of the Fourth Eastern and Southern Africa Regional Maize Conference, Harare, Zimbabwe CIMMYT, Mexico City, Mexico*, pp. 202-204.
- Blinkova, O., Victoria, J., Li, Y., Keele, B. F., Sanz, C., Ndjango, J.-B. N., Peeters, M., Travis, D., Lonsdorf, E. V., Wilson, M. L., Pusey, A. E., Hahn, B. H. & Delwart, E. L. (2010).** Novel circular DNA viruses in stool samples of wild-living chimpanzees. *Journal of General Virology* **91**, 74-86.
- Bock, K., Guthrie, E. & Woods, R. (1974).** Purification of maize streak virus and its relationship to viruses associated with streak diseases of sugar cane and *Panicum maximum*. *Annals of applied biology* **77**, 289-296.
- Bosque-Pérez, N. A. (2000).** Eight decades of maize streak virus research. *Virus Research* **71**, 107-121.
- Boulton, M. I. (2002).** Functions and interactions of mastrevirus gene products. *Physiological and Molecular Plant Pathology* **60**, 243-255.
- Boulton, M. I., Pallaghy, C. K., Chatani, M., MacFarlane, S. & Davies, J. W. (1993).** Replication of Maize Streak Virus Mutants in Maize Protoplasts: Evidence for a Movement Protein. *Virology* **192**, 85-93.
- Boulton, M. I., Steinkellner, H., Donson, J., Markham, P. G., King, D. I. & Davies, J. W. (1989).** Mutational analysis of the virion-sense genes of maize streak virus. *Journal of General Virology* **70**, 2309-2323.
- Briddon, R., Pinner, M., Stanley, J. & Markham, P. (1990).** Geminivirus coat protein gene replacement alters insect specificity. *Virology* **177**, 85-94.
- Briddon, R. W., Bedford, I. D., Tsai, J. H. & Markham, P. G. (1996).** Analysis of the Nucleotide Sequence of the Treehopper-Transmitted Geminivirus, Tomato Pseudo-Curly Top Virus, Suggests a Recombinant Origin. *Virology* **219**, 387-394.
- Briddon, R. W., Heydarnejad, J., Khosrowfar, F., Massumi, H., Martin, D. P. & Varsani, A. (2010a).** Turnip curly top virus, a highly divergent geminivirus infecting turnip in Iran. *Virus Research* **152**, 169-175.
- Briddon, R. W., Lunness, P., Chamberlin, L. C. L., Pinner, M. S., Brundish, H. & Markham, P. G. (1992).** The nucleotide sequence of an infectious insect-transmissible clone of the geminivirus *Panicum streak virus*. *Journal of General Virology* **73**, 1041-1047.
- Briddon, R. W., Martin, D. P., Owor, B. E., Donaldson, L., Markham, P. G., Greber, R. S. & Varsani, A. (2010b).** A novel species of mastrevirus (family Geminiviridae) isolated from *Digitaria didactyla* grass from Australia. *Arch Virol* **155**, 1529-1534.
- Briddon, R. W., Stenger, D. C., Bedford, I. D., Stanley, J., Izadpanah, K. & Markham, P. G. (1998).** Comparison of a beet curly top virus isolate originating from the old world with those from the new world. *European journal of plant pathology* **104**, 77-84.
- Briddon, R. W., Watts, J., Markham, P. G. & Stanley, J. (1989).** The coat protein of beet curly top virus is essential for infectivity. *Virology* **172**, 628-633.
- Brown, J. K., Fauquet, C. M., Briddon, R. W., Zerbini, M., Moriones, E. & Navas-Castillo, J. (2011).** *Virus taxonomy: classification and nomenclature of viruses: Ninth Report of*

- the International Committee on Taxonomy of Viruses*: Elsevier Academic Press, San Diego.
- Candresse, T., Filloux, D., Muhire, B., Julian, C., Galzi, S., Fort, G., Bernardo, P., Daugrois, J.-H., Fernandez, E., Martin, D. P., Varsani, A. & Roumagnac, P. (2014).** Appearances Can Be Deceptive: Revealing a Hidden Viral Infection with Deep Sequencing in a Plant Quarantine Context. *PLoS ONE* **9**, e102945.
- Cantalupo, P. G., Calgua, B., Zhao, G., Hundesa, A., Wier, A. D., Katz, J. P., Grabe, M., Hendrix, R. W., Girones, R., Wang, D. & Pipas, J. M. (2011).** Raw Sewage Harbors Diverse Viral Populations. *mBio* **2**, e00180-00111.
- Casado, C. G., Javier Ortiz, G., Padron, E., Bean, S. J., McKenna, R., Agbandje-McKenna, M. & Boulton, M. I. (2004).** Isolation and characterization of subgenomic DNAs encapsidated in “single” $T=1$ isometric particles of *Maize streak virus*. *Virology* **323**, 164-171.
- Castellano, M., Sanz-Burgos, A. P. & Gutiérrez, C. (1999).** Initiation of DNA replication in a eukaryotic rolling-circle replicon: identification of multiple DNA-protein complexes at the geminivirus origin. *Journal of molecular biology* **290**, 639-652.
- Caulfield, J. L., Wishnok, J. S. & Tannenbaum, S. R. (1998).** Nitric oxide-induced deamination of cytosine and guanine in deoxynucleosides and oligonucleotides. *Journal of Biological Chemistry* **273**, 12689-12695.
- Cenchrus, C. & Coix, D. (1991).** Characterization of maize streak virus isolates using monoclonal and polyclonal antibodies and by transmission to a few hosts. *Plant disease*, **27**.
- Chatani, M., Matsumoto, Y., Mizuta, H., Ikegami, M., Boulton, M. I. & Davies, J. W. (1991).** The nucleotide sequence and genome structure of the geminivirus miscanthus streak virus. *The Journal of general virology* **72** (10), 2325-2331.
- Chen, L.-F. & Gilbertson, R. L. (2008).** Curtovirus–Cucurbit Interaction: Acquisition Host Plays a Role in Leafhopper Transmission in a Host-Dependent Manner. *Phytopathology* **99**, 101-108.
- Chen, L.-F., Vivoda, E. & Gilbertson, R. (2011).** Genetic diversity in curtoviruses: a highly divergent strain of Beet mild curly top virus associated with an outbreak of curly top disease in pepper in Mexico. *Arch Virol* **156**, 547-555.
- Collin, S., Fernández-lobato, M., Gooding, P. S., Mullineaux, P. M. & Fenoll, C. (1996).** The two nonstructural proteins from wheat dwarf virus involved in viral gene expression and replication are retinoblastoma-binding proteins. *Virology* **219**, 324-329.
- Creamer, R., Hubble, H. & Lewis, A. (2005).** Curtovirus infection of chile pepper in New Mexico. *Plant disease* **89**, 480-486.
- Dabrowski, Z. (1987).** Two new species of Cicadulina China (Hemiptera: Euscelidae) from West Africa. *Bulletin of entomological research* **77**, 53-56.
- Dayaram, A., Opong, A., Jäschke, A., Hadfield, J., Baschiera, M., Dobson, R. C. J., Offei, S. K., Shepherd, D. N., Martin, D. P. & Varsani, A. (2012).** Molecular characterisation of a novel cassava associated circular ssDNA virus. *Virus Research* **166**, 130-135.
- Dayaram, A., Potter, K. A., Moline, A. B., Rosenstein, D. D., Marinov, M., Thomas, J. E., Breitbart, M., Rosario, K., Argüello-Astorga, G. R. & Varsani, A. (2013).** High global diversity of cycloviruses amongst dragonflies. *Journal of General Virology* **94**, 1827-1840.

- De Bruyn, A., Villemot, J., Lefeuvre, P., Villar, E., Hoareau, M., Harimalala, M., Abdoul-Karime, A. L., Abdou-Chakour, C., Reynaud, B. & Harkins, G. W. (2012).** East African cassava mosaic-like viruses from Africa to Indian ocean islands: molecular diversity, evolutionary history and geographical dissemination of a bipartite begomovirus. *BMC evolutionary biology* **12**, 228.
- de Jong, M. D., Van Kinh, N., Trung, N. V., Taylor, W., Wertheim, H. F., van der Ende, A., van der Hoek, L., Canuti, M., Crusat, M. & Sona, S. (2014).** Limited geographic distribution of the novel cyclovirus CyCV-VN. *Scientific reports* **4**.
- Dekker, E. L., Woolston, C. J., Xue, Y., Cox, B. & Mullineaux, P. M. (1991).** Transcript mapping reveals different expression strategies for the bicistronic RNAs of the geminivirus wheat dwarf virus. *Nucleic Acids Research* **19**, 4075-4081.
- Dickinson, V. J., Halder, J. & Woolston, C. J. (1996).** The Product of Maize Streak Virus ORF V1 Is Associated with Secondary Plasmodesmata and Is First Detected with the Onset of Viral Lesions. *Virology* **220**, 51-59.
- Donson, J., Accotto, G. P., Boulton, M. I., Mullineaux, P. M. & Davies, J. W. (1987).** The nucleotide sequence of a geminivirus from *Digitaria sanguinalis*. *Virology* **161**, 160-169.
- Donson, J., Morris-Krsinich, B., Mullineaux, P., Boulton, M. & Davies, J. (1984).** A putative primer for second-strand DNA synthesis of maize streak virus is virion-associated. *The EMBO journal* **3**, 3069.
- Du, Z., Tang, Y., Zhang, S., She, X., Lan, G., Varsani, A. & He, Z. (2014).** Identification and molecular characterization of a single-stranded circular DNA virus with similarities to *Sclerotinia sclerotiorum* hypovirulence-associated DNA virus 1. *Arch Virol* **159**, 1527-1531.
- Duffy, S. & Holmes, E. C. (2008).** Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. *Journal of Virology* **82**, 957-965.
- Duffy, S. & Holmes, E. C. (2009).** Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *Journal of general virology* **90**, 1539-1547.
- Duffy, S., Shackelton, L. A. & Holmes, E. C. (2008).** Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics* **9**, 267-276.
- Ekzayez, A., Kumari, S. & Ismail, I. (2011).** First report of wheat dwarf virus and its vector (*Psammotettix provincialis*) affecting wheat and barley crops in Syria. *Plant Disease* **95**, 76-76.
- Erdmann, J. B., Shepherd, D. N., Martin, D. P., Varsani, A., Rybicki, E. P. & Jeske, H. (2010).** Replicative intermediates of maize streak virus found during leaf development. *Journal of general virology* **91**, 1077-1081.
- Farzadfar, S., Pourrahim, R., Golnaraghi, A. R. & Ahoonmanesh, A. (2008).** PCR detection and partial molecular characterization of Chickpea chlorotic dwarf virus in naturally infected sugar beet plants in Iran. *Journal of Plant Pathology* **90**, 247-251.
- Fauquet, C., Briddon, R., Brown, J., Moriones, E., Stanley, J., Zerbini, M. & Zhou, X. (2008).** Geminivirus strain demarcation and nomenclature. *Arch Virol* **153**, 783-821.
- Fauquet, C. M. & Stanley, J. (2003).** Geminivirus classification and nomenclature: progress and problems. *Annals of Applied Biology* **142**, 165-189.

- Fenoll, C., Schwarz, J. J., Black, D. M., Schneider, M. & Howell, S. H. (1990).** The intergenic region of maize streak virus contains a GC-rich element that activates rightward transcription and binds maize nuclear factors. *Plant molecular biology* **15**, 865-877.
- Firth, C., Charleston, M. A., Duffy, S., Shapiro, B. & Holmes, E. C. (2009).** Insights into the Evolutionary History of an Emerging Livestock Pathogen: Porcine Circovirus 2. *Journal of Virology* **83**, 12813-12821.
- Fletcher, M. (2009).** Identification keys and checklists for the leafhoppers, planthoppers and their relatives occurring in Australia and neighbouring areas (Hemiptera: Auchenorrhyncha). URL <http://www.agric.nsw.gov.au/Hort/ascu/start.htm> [accessed on 1 July 2009].
- Fondong, V. N. (2013).** Geminivirus protein structure and function. *Molecular plant pathology* **14**, 635-649.
- Fontes, E., Luckow, V. A. & Hanley-Bowdoin, L. (1992).** A geminivirus replication protein is a sequence-specific DNA binding protein. *The Plant Cell Online* **4**, 597-608.
- Forterre, P. (1992).** New hypotheses about the origins of viruses, prokaryotes and eukaryotes. *Frontiers of Life, Van TT, Mounolou JK, Shneider J, Mc Kay C editor Editions Frontieres Gif-sur-Yvette Cedex, France*, 221-234.
- Forterre, P. (2006).** The origin of viruses and their possible roles in major evolutionary transitions. *Virus research* **117**, 5-16.
- Gafni, Y. & Epel, B. L. (2002).** The role of host and viral proteins in intra- and inter-cellular trafficking of geminiviruses. *Physiological and Molecular Plant Pathology* **60**, 231-241.
- Ge, X., Li, J., Peng, C., Wu, L., Yang, X., Wu, Y., Zhang, Y. & Shi, Z. (2011).** Genetic diversity of novel circular ssDNA viruses in bats in China. *Journal of General Virology* **92**, 2646-2653.
- Geering, A. D. W., Thomas, J. E., Holton, T., Hadfield, J. & Varsani, A. (2011).** Paspalum striate mosaic virus: an Australian mastrevirus from Paspalum dilatatum. *Arch Virol* **157**, 193-197.
- Geslin, C., Le Romancer, M., Gaillard, M., Erauso, G. & Prieur, D. (2003).** Observation of virus-like particles in high temperature enrichment cultures from deep-sea hydrothermal vents. *Research in microbiology* **154**, 303-307.
- Gharouni Kardani, S., Heydarnejad, J., Zakiaghl, M., Mehrvar, M., Kraberger, S. & Varsani, A. (2013).** Diversity of Beet curly top Iran virus isolated from different hosts in Iran. *Virus Genes* **46**, 571-575.
- Gorbalenya, A. E. & Koonin, E. V. (1993).** Helicases: amino acid sequence comparisons and structure-function relationships. *Current Opinion in Structural Biology* **3**, 419-429.
- Gorbalenya, A. E., Koonin, E. V. & Wolf, Y. I. (1990).** A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. *FEBS letters* **262**, 145-148.
- Greber, R. (1989).** Biological characteristics of grass geminiviruses from eastern Australia. *Annals of Applied Biology* **114**, 471-480.
- Grigoras, I., Timchenko, T., Grande-Pérez, A., Katul, L., Vetten, H.-J. & Gronenborn, B. (2010).** High variability and rapid evolution of a nanovirus. *Journal of virology* **84**, 9105-9117.
- Gröning, B. R., Frischmuth, T. & Jeske, H. (1990).** Replicative form DNA of abutilon mosaic virus is present in plastids. *Mol Gen Genet* **220**, 485-488.

- Gutierrez, C. (1999).** Geminivirus DNA replication. *Cellular and Molecular Life Sciences* **56**, 313-329.
- Gutierrez, C., Ramirez-Parra, E., Mar Castellano, M., Sanz-Burgos, A. P., Luque, A. & Missich, R. (2004).** Geminivirus DNA replication and cell cycle interactions. *Veterinary microbiology* **98**, 111-119.
- Gutiérrez, C., Suárez-López, P., Ramírez-Parra, E., Sanz-Burgos, A., Pönninger, J. & Xie, Q. (1995).** DNA bending as a potential regulatory cis-acting element of the geminivirus intergenic region. *Agronomie* **15**, 415-420.
- Hadfield, J., Martin, D., Stainton, D., Krabberger, S., Owor, B., Shepherd, D., Lakay, F., Markham, P., Greber, R., Briddon, R. & Varsani, A. (2011).** Bromus catharticus striate mosaic virus: a new mastrevirus infecting *Bromus catharticus* from Australia. *Arch Virol* **156**, 335-341.
- Hadfield, J., Thomas, J. E., Schwinghamer, M. W., Krabberger, S., Stainton, D., Dayaram, A., Parry, J. N., Pande, D., Martin, D. P. & Varsani, A. (2012).** Molecular characterisation of dicot-infecting mastreviruses from Australia. *Virus Research* **166**, 13-22.
- Haible, D., Kober, S. & Jeske, H. (2006).** Rolling circle amplification revolutionizes diagnosis and genomics of geminiviruses. *Journal of virological methods* **135**, 9-16.
- Hallan, V. & Gafni, Y. (2001).** Tomato yellow leaf curl virus (TYLCV) capsid protein (CP) subunit interactions: implications for viral assembly. *Arch Virol* **146**, 1765-1773.
- Halley-Stott, R. P., Tanzer, F., Martin, D. P. & Rybicki, E. P. (2007).** The complete nucleotide sequence of a mild strain of Bean yellow dwarf virus. *Arch Virol* **152**, 1237-1240.
- Hanley-Bowdoin, L., Bejarano, E. R., Robertson, D. & Mansoor, S. (2013).** Geminiviruses: masters at redirecting and reprogramming plant processes. *Nature Reviews Microbiology* **11**, 777-788.
- Harkins, G. W., Delpont, W., Duffy, S., Wood, N., Monjane, A. L., Owor, B. E., Donaldson, L., Saumtally, S., Triton, G., Briddon, R. W., Shepherd, D. N., Rybicki, E. P., Martin, D. P. & Varsani, A. (2009a).** Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts. *Virology Journal* **6**.
- Harkins, G. W., Martin, D. P., Christoffels, A. & Varsani, A. (2014).** Towards inferring the global movement of beak and feather disease virus. *Virology* **450**, 24-33.
- Harkins, G. W., Martin, D. P., Duffy, S., Monjane, A. L., Shepherd, D. N., Windram, O. P., Owor, B. E., Donaldson, L., van Antwerpen, T., Sayed, R. A., Flett, B., Ramusi, M., Rybicki, E. P., Peterschmitt, M. & Varsani, A. (2009b).** Dating the origins of the maize-adapted strain of maize streak virus, MSV-A. *Journal of General Virology* **90**, 3066-3074.
- Hayes, R., MacDonald, H., Coutts, R. & Buck, K. (1988).** Priming of complementary DNA synthesis in vitro by small DNA molecules tightly bound to virion DNA of wheat dwarf virus. *Journal of general virology* **69**, 1345-1350.
- Hendrix, R. W., Lawrence, J. G., Hatfull, G. F. & Casjens, S. (2000).** The origins and ongoing evolution of viruses. *Trends in microbiology* **8**, 504-508.
- Hernández-Zepeda, C., Varsani, A. & Brown, J. (2013).** Intergeneric recombination between a new, spinach-infecting curtovirus and a new geminivirus belonging to the genus Becurtovirus: first New World exemplar. *Arch Virol* **158**, 2245-2254.

- Heydarnejad, J., Keyvani, N., Razavinejad, S., Massumi, H. & Varsani, A. (2013).** Fulfilling Koch's postulates for beet curly top Iran virus and proposal for consideration of new genus in the family Geminiviridae. *Arch Virol* **158**, 435-443.
- Heyraud, F., Matzeit, V., Schaefer, S., Schell, J. & Gronenborn, B. (1993).** The conserved nonanucleotide motif of the geminivirus stem-loop sequence promotes replicational release of virus molecules from redundant copies. *Biochimie* **75**, 605-615.
- Hipp, K., Rau, P., Schäfer, B., Gronenborn, B. & Jeske, H. (2014).** The RXL motif of the African cassava mosaic virus Rep protein is necessary for rereplication of yeast DNA and viral infection in plants. *Virology* **462**, 189-198.
- Hofer, J., Dekker, E. L., Reynolds, H. V., Woolston, C. J., Cox, B. S. & Mullineaux, P. M. (1992).** Coordinate regulation of replication and virion sense gene expression in wheat dwarf virus. *The Plant Cell Online* **4**, 213-223.
- Höfer, P., Bedford, I. D., Markham, P. G., Jeske, H. & Frischmuth, T. (1997).** Coat protein gene replacement results in whitefly transmission of an insect nontransmissible geminivirus isolate. *Virology* **236**, 288-295.
- Höhnle, M., Höfer, P., Bedford, I. D., Briddon, R. W., Markham, P. G. & Frischmuth, T. (2001).** Exchange of Three Amino Acids in the Coat Protein Results in Efficient Whitefly Transmission of a Nontransmissible *Abutilon Mosaic Virus* Isolate. *Virology* **290**, 164-171.
- Hormuzdi, S. G. & Bisaro, D. M. (1993).** Genetic analysis of beet curly top virus: evidence for three virion sense genes involved in movement and regulation of single-and double-stranded DNA levels. *Virology* **193**, 900-909.
- Horn, N. M., Reddy, S. V. & Reddy, D. V. R. (1994).** Virus-vector relationships of chickpea chlorotic dwarf geminivirus and the leafhopper *Orosius orientalis* (Hemiptera: Cicadellidae). *Annals of Applied Biology* **124**, 441-450.
- Horn, N. M., Reddy, S. V., Roberts, I. M. & Reddy, D. V. R. (1993).** Chickpea chlorotic dwarf virus, a new leafhopper-transmitted geminivirus of chickpea in India. *Annals of Applied Biology* **122**, 467-479.
- Horváth, G., Pettkó-Szandtner, A., Nikovics, K., Bilgin, M., Boulton, M., Davies, J., Gutiérrez, C. & Dudits, D. (1998).** Prediction of functional regions of the maize streak virus replication-associated proteins by protein-protein interaction analysis. *Plant Molecular Biology* **38**, 699-712.
- Hughes, F., Rybicki, E. & Kirby, R. (1993).** Complete nucleotide sequence of sugarcane streak Monogeminivirus. *Arch Virol* **132**, 171-182.
- Ilyina, T. V. & Koonin, E. V. (1992).** Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaeobacteria. *Nucleic Acids Research* **20**, 3279-3285.
- Inamdar, N. M., Zhang, X.-Y., Brough, C. L., Gardiner, W. E., Bisaro, D. M. & Ehrlich, M. (1992).** Transfection of heteroduplexes containing uracil· guanine or thymine· guanine mispairs into plant cells. *Plant molecular biology* **20**, 123-131.
- Inoue-Nagata, A. K., Albuquerque, L. C., Rocha, W. B. & Nagata, T. (2004).** A simple method for cloning the complete begomovirus genome using the bacteriophage ϕ 29 DNA polymerase. *Journal of virological methods* **116**, 209-211.
- Isnard, M., Granier, M., Frutos, R., Reynaud, B. & Peterschmitt, M. (1998).** Quasispecies nature of three maize streak virus isolates obtained through different modes of selection

- from a population used to assess response to infection of maize cultivars. *Journal of General Virology* **79**, 3091-3099.
- Jeske, H., Lütgemeier, M. & Preiß, W. (2001).** DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. *The EMBO journal* **20**, 6158-6167.
- Jørgensen, T. S., Xu, Z., Hansen, M. A., Sørensen, S. J. & Hansen, L. H. (2014).** Hundreds of Circular Novel Plasmids and DNA Elements Identified in a Rat Cecum Metamobilome. *PloS one* **9**, e87924.
- Kacprzak, M., Neczaj, E. & Okoniewska, E. (2005).** The comparative mycological analysis of wastewater and sewage sludges from selected wastewater treatment plants. *Desalination* **185**, 363-370.
- Kammann, M., Schalk, H.-J., Matzeit, V., Schaefer, S., Schell, J. & Gronenborn, B. (1991).** DNA replication of wheat dwarf virus, a geminivirus, requires two *cis*-acting signals. *Virology* **184**, 786-790.
- Kenton, A., Khashoggi, A., Parokonny, A., Bennett, M. & Lichtenstein, C. (1995).** Chromosomal location of endogenous geminivirus-related DNA sequences in *Nicotiana tabacum* L. *Chromosome Research* **3**, 346-350.
- Kepner Jr, R. L., Wharton Jr, R. A. & Suttle, C. A. (1998).** Viruses in Antarctic lakes. *Limnology and Oceanography* **43**, 1754-1761.
- Kim, H. K., Park, S. J., Nguyen, V. G., Song, D. S., Moon, H. J., Kang, B. K. & Park, B. K. (2012).** Identification of a novel single-stranded, circular DNA virus from bovine stool. *Journal of General Virology* **93**, 635-639.
- Kim, K.-H., Chang, H.-W., Nam, Y.-D., Roh, S. W., Kim, M.-S., Sung, Y., Jeon, C. O., Oh, H.-M. & Bae, J.-W. (2008).** Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Applied and environmental microbiology* **74**, 5975-5985.
- King, A. M., Adams, M. J., Lefkowitz, E. J. & Carstens, E. B. (2012).** *Virus taxonomy: classification and nomenclature of viruses: Ninth Report of the International Committee on Taxonomy of Viruses*: Elsevier.
- Klute, K. A., Nadler, S. A. & Stenger, D. C. (1996).** Horseradish curly top virus is a distinct subgroup II geminivirus species with rep and C4 genes derived from a subgroup III ancestor. *Journal of general virology* **77**, 1369-1378.
- Köklü, G., Ramsell, J. N. & Kvarnheden, A. (2007).** The complete genome sequence for a Turkish isolate of Wheat dwarf virus (WDV) from barley confirms the presence of two distinct WDV strains. *Virus Genes* **34**, 359-366.
- Koonin, E. V. & Ilyina, T. V. (1992).** Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. *The Journal of General Virology* **73** 2763-2766.
- Koonin, E. V., Senkevich, T. G. & Dolja, V. V. (2006).** The ancient Virus World and evolution of cells. *Biol Direct* **1**, 29.
- Kotlizky, G., Boulton, M. I., Pitaksutheepong, C., Davies, J. W. & Epel, B. L. (2000).** Intracellular and Intercellular Movement of Maize Streak Geminivirus V1 and V2 Proteins Transiently Expressed as Green Fluorescent Protein Fusions. *Virology* **274**, 32-38.
- Kraberger, S., Harkins, G. W., Kumari, S. G., Thomas, J. E., Schwinghamer, M. W., Sharman, M., Collings, D. A., Briddon, R. W., Martin, D. P. & Varsani, A. (2013a).** Evidence that dicot-infecting mastreviruses are particularly prone to inter-species

- recombination and have likely been circulating in Australia for longer than in Africa and the Middle East. *Virology* **444**, 282-291.
- Krabberger, S., Stainton, D., Dayaram, A., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2013b).** Discovery of Sclerotinia sclerotiorum Hypovirulence-Associated Virus-1 in Urban River Sediments of Heathcote and Styx Rivers in Christchurch City, New Zealand. *Genome Announcements* **1**, e00559-00513.
- Krabberger, S., Thomas, J. E., Geering, A. D. W., Dayaram, A., Stainton, D., Hadfield, J., Walters, M., Parmenter, K. S., van Brunschot, S., Collings, D. A., Martin, D. P. & Varsani, A. (2012).** Australian monocot-infecting mastrevirus diversity rivals that in Africa. *Virus Research* **169**, 127-136.
- Krenz, B., Thompson, J. R., Fuchs, M. & Perry, K. L. (2012).** Complete Genome Sequence of a New Circular DNA Virus from Grapevine. *Journal of Virology* **86**, 7715.
- Kreuze, J. F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I. & Simon, R. (2009).** Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. *Virology* **388**, 1-7.
- Krupovic, M., Prangishvili, D., Hendrix, R. W. & Bamford, D. H. (2011).** Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. *Microbiology and Molecular Biology Reviews* **75**, 610-635.
- Krupovic, M., Ravantti, J. J. & Bamford, D. H. (2009).** Geminiviruses: a tale of a plasmid becoming a virus. *BMC evolutionary biology* **9**, 112.
- Kumar, J., Kumar, J., Singh, S. P. & Tuli, R. (2014).** Association of satellites with a mastrevirus in natural infection: complexity of Wheat dwarf India virus disease. *Journal of Virology* **88**, 7093-7104.
- Kumar, J., Singh, S. P., Kumar, J. & Tuli, R. (2012).** A novel mastrevirus infecting wheat in India. *Arch Virol* **157**, 2031-2034.
- Kumari, S. G., Makkouk, K. M. & Attar, N. (2006).** An improved antiserum for sensitive serologic detection of Chickpea chlorotic dwarf virus. *Journal of Phytopathology* **154**, 129-133.
- Kumari, S. G., Makkouk, K. M., Attar, N., Ghulam, W. & Lesemann, D. E. (2004).** First Report of Chickpea chlorotic dwarf virus infecting spring chickpea in Syria. *Plant Disease* **88**, 424-424.
- Kumari, S. G., Makkouk, K. M., Loh, M. H., Negassi, K., Tsegay, S., Kidane, R., Kibret, A. & Tesfatsion, Y. (2008).** Viral diseases affecting chickpea crops in Eritrea. *Phytopathologia Mediterranea* **47**, 42-49.
- Kvarnheden, A., Lindblad, M., Lindsten, K. & Valkonen, J. (2002).** Genetic diversity of Wheat dwarf virus. *Arch Virol* **147**, 205-216.
- Labonté, J. M. & Suttle, C. A. (2013).** Previously unknown and highly divergent ssDNA viruses populate the oceans. *The ISME journal* **7**, 2169-2177.
- Lam, N., Creamer, R., Rascon, J. & Belfon, R. (2009).** Characterization of a new curtovirus, pepper yellow dwarf virus, from chile pepper and distribution in weed hosts in New Mexico. *Arch Virol* **154**, 429-436.
- Lapierre, H. & Signoret, P. A. (2004).** *Viruses and virus diseases of Poaceae (Gramineae)*: Editions Quae.

- Laufs, J., Schumacher, S., Geisler, N., Jupin, I. & Gronenborn, B. (1995a).** Identification of the nicking tyrosine of geminivirus Rep protein. *FEBS letters* **377**, 258-262.
- Laufs, J., Traut, W., Heyraud, F., Matzeit, V., Rogers, S. G., Schell, J. & Gronenborn, B. (1995b).** In vitro cleavage and joining at the viral origin of replication by the replication initiator protein of tomato yellow leaf curl virus. *Proceedings of the National Academy of Sciences* **92**, 3879-3883.
- Lawry, R., Martin, D., Shepherd, D., van Antwerpen, T. & Varsani, A. (2009).** A novel sugarcane-infecting mastrevirus from South Africa. *Arch Virol* **154**, 1699-1703.
- Laybourn Parry, J., Hofer, J. S. & Sommaruga, R. (2001).** Viruses in the plankton of freshwater and saline Antarctic lakes. *Freshwater Biology* **46**, 1279-1287.
- Lazarowitz, S. G., Pinder, A. J., Damsteegt, V. D. & Rogers, S. G. (1989).** Maize streak virus genes essential for systemic spread and symptom development. *The EMBO journal* **8**, 1023.
- Lefevre, P., Harkins, G. W., Lett, J.-M., Briddon, R. W., Chase, M. W., Moury, B. & Martin, D. P. (2011).** Evolutionary time-scale of the begomoviruses: evidence from integrated sequences in the Nicotiana genome. *PLoS One* **6**, e19193.
- Lefevre, P., Lett, J.-M., Varsani, A. & Martin, D. (2009).** Widely conserved recombination patterns among single-stranded DNA viruses. *Journal of virology* **83**, 2697-2707.
- Lefevre, P., Martin, D. P., Harkins, G., Lemey, P., Gray, A. J., Meredith, S., Lakay, F., Monjane, A., Lett, J.-M. & Varsani, A. (2010).** The spread of Tomato yellow leaf curl virus from the Middle East to the world. *PLoS Pathogens* **6**, e1001164.
- Lett, J.-M., Granier, M., Hippolyte, I., Grondin, M., Royer, M., Blanc, S., Reynaud, B. & Peterschmitt, M. (2002).** Spatial and temporal distribution of geminiviruses in leafhoppers of the genus Cicadulina monitored by conventional and quantitative polymerase chain reaction. *Phytopathology* **92**, 65-74.
- Li, L., Kapoor, A., Slikas, B., Bamidele, O. S., Wang, C., Shaukat, S., Masroor, M. A., Wilson, M. L., Ndjanga, J.-B. N., Peeters, M., Gross-Camp, N. D., Muller, M. N., Hahn, B. H., Wolfe, N. D., Triki, H., Bartkus, J., Zaidi, S. Z. & Delwart, E. (2010).** Multiple Diverse Circoviruses Infect Farm Animals and Are Commonly Found in Human and Chimpanzee Feces. *Journal of Virology* **84**, 1674-1682.
- Lim, K. Y., Matyášek, R., Lichtenstein, C. P. & Leitch, A. R. (2000).** Molecular cytogenetic analyses and phylogenetic studies in the Nicotiana section Tomentosae. *Chromosoma* **109**, 245-258.
- Liu, H., Boulton, M. I. & Davies, J. W. (1997a).** Maize streak virus coat protein binds single- and double-stranded DNA in vitro. *Journal of General Virology* **78**, 1265-1270.
- Liu, H., Boulton, M. I., Oparka, K. J. & Davies, J. W. (2001).** Interaction of the movement and coat proteins of Maize streak virus: Implications for the transport of viral DNA. *Journal of General Virology* **82**, 35-44.
- Liu, H., Boulton, M. I., Thomas, C. L., Prior, D. A. M., Oparka, K. J. & Davies, J. W. (1999a).** Maize streak virus coat protein is karyophyllic and facilitates nuclear transport of viral DNA. *Molecular Plant-Microbe Interactions* **12**, 894-900.
- Liu, H., Fu, Y., Li, B., Yu, X., Xie, J., Cheng, J., Ghabrial, S. A., Li, G., Yi, X. & Jiang, D. (2011).** Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC Evolutionary Biology* **11**, 276.

- Liu, L., Davies, J. W. & Stanley, J. (1998).** Mutational analysis of bean yellow dwarf virus, a geminivirus of the genus Mastrevirus that is adapted to dicotyledonous plants. *Journal of general virology* **79**, 2265-2274.
- Liu, L., Saunders, K., Thomas, C. L., Davies, J. W. & Stanley, J. (1999b).** Bean yellow dwarf virus RepA, but not Rep, binds to maize retinoblastoma protein, and the virus tolerates mutations in the consensus binding motif. *Virology* **256**, 270-279.
- Liu, L., van Tonder, T., Pietersen, G., Davies, J. W. & Stanley, J. (1997b).** Molecular characterization of a subgroup I geminivirus from a legume in South Africa. *Journal of General Virology* **78**, 2113-2117.
- Loconsole, G., Saldarelli, P., Doddapaneni, H., Savino, V., Martelli, G. P. & Saponari, M. (2012).** Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease, a new member in the family Geminiviridae. *Virology* **432**, 162-172.
- Londoño, A., Riego-Ruiz, L. & Argüello-Astorga, G. (2010).** DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. *Arch Virol* **155**, 1033-1046.
- López-Bueno, A., Tamames, J., Velázquez, D., Moya, A., Quesada, A. & Alcamí, A. (2009).** High Diversity of the Viral Community from an Antarctic Lake. *Science* **326**, 858-861.
- Lyttle, D. & Guy, P. (2004).** First record of Geminiviruses in New Zealand: Abutilon mosaic virus and Honeysuckle yellow vein virus. *Australasian Plant Pathology* **33**, 321-322.
- MacDowell, S., Macdonald, H., Hamilton, W., Coutts, R. & Buck, K. (1985).** The nucleotide sequence of cloned wheat dwarf virus DNA. *The EMBO journal* **4**, 2173.
- Makkouk, K. M., Hamed, A. A., Hussein, M. & Kumari, S. G. (2003a).** First report of Faba bean necrotic yellows virus (FBNYV) infecting chickpea (*Cicer arietinum*) and faba bean (*Vicia faba*) crops in Sudan. *Plant Pathology* **52**, 412-412.
- Makkouk, K. M., Rizkallah, L., Kumari, S. G., Zaki, M. & Enein, R. A. (2003b).** First record of Chickpea chlorotic dwarf virus (CpCDV) affecting faba bean (*Vicia faba*) crops in Egypt. *Plant Pathology* **52**, 413-413.
- Mansoor, S., Briddon, R. W., Zafar, Y. & Stanley, J. (2003).** Geminivirus disease complexes: an emerging threat. *Trends in Plant Science* **8**, 128-134.
- Manzoor, M., Ilyas, M., Shafiq, M., Haider, M., Shahid, A. & Briddon, R. (2014).** A distinct strain of chickpea chlorotic dwarf virus (genus Mastrevirus, family Geminiviridae) identified in cotton plants affected by leaf curl disease. *Arch Virol* **159** (5), 1217-1221.
- Mardis, E. R. (2008).** The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24**, 133-141.
- Martin, D. P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P. & Varsani, A. (2011a).** Recombination in Eukaryotic Single Stranded DNA Viruses. *Viruses* **3**, 1699-1738.
- Martin, D. P., Briddon, R. W. & Varsani, A. (2011b).** Recombination patterns in dicot-infecting mastreviruses mirror those found in monocot-infecting mastreviruses. *Arch Virol* **156**, 1463-1469.
- Martin, D. P., Linderme, D., Lefeuvre, P., Shepherd, D. N. & Varsani, A. (2011c).** Eragrostis minor streak virus: an Asian streak virus in Africa. *Arch Virol* **156**, 1299-1303.
- Martin, D. P., Willment, J. A., Billharz, R., Velders, R., Odhiambo, B., Njuguna, J., James, D. & Rybicki, E. P. (2001).** Sequence diversity and virulence in Zea mays of Maize streak virus isolates. *Virology* **288**, 247-255.

- Massart, S., Olmos, A., Jijakli, H. & Candresse, T. (2014).** Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus research* **188**, 90-96.
- Mauricio-Castillo, J., Torres-Herrera, S., Cárdenas-Conejo, Y., Pastor-Palacios, G., Méndez-Lozano, J. & Argüello-Astorga, G. (2014).** A novel begomovirus isolated from sida contains putative cis-and trans-acting replication specificity determinants that have evolved independently in several geographical lineages. *Arch Virol*, 1-12.
- McDaniel, L. D., Rosario, K., Breitbart, M. & Paul, J. H. (2013).** Comparative metagenomics: Natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. *Environmental microbiology* **16**, 570-585.
- McGivern, D. R., Findlay, K. C., Montague, N. P. & Boulton, M. I. (2005).** An intact RBR-binding motif is not required for infectivity of Maize streak virus in cereals, but is required for invasion of mesophyll cells. *Journal of general virology* **86**, 797-801.
- Metzker, M. L. (2010).** Sequencing technologies—the next generation. *Nature Reviews Genetics* **11**, 31-46.
- Monjane, A., van der Walt, E., Varsani, A., Rybicki, E. & Martin, D. (2011a).** Recombination hotspots and host susceptibility modulate the adaptive value of recombination during maize streak virus evolution. *BMC Evolutionary Biology* **11**, 350.
- Monjane, A. L., Harkins, G. W., Martin, D. P., Lemey, P., Lefevre, P., Shepherd, D. N., Oluwafemi, S., Simuyandi, M., Zinga, I., Komba, E. K., Lakoutene, D. P., Mandakombo, N., Mboukoulida, J., Semballa, S., Tagne, A., Tiendrebeogo, F., Erdmann, J. B., van Antwerpen, T., Owor, B. E., Flett, B., Ramusi, M., Windram, O. P., Syed, R., Lett, J. M., Briddon, R. W., Markham, P. G., Rybicki, E. P. & Varsani, A. (2011b).** Reconstructing the history of maize streak virus strain a dispersal to reveal diversification hot spots and its origin in southern Africa. *Journal of Virology* **85**, 9623-9636.
- Monjane, A. L., Martin, D. P., Lakay, F., Muhire, B. M., Pande, D., Varsani, A., Harkins, G., Shepherd, D. N. & Rybicki, E. P. (2014).** Extensive recombination—induced disruption of genetic interactions is highly deleterious but can be partially reversed by small numbers of secondary-recombination events. *Journal of Virology*, JVI-00709.
- Monjane, A. L., Pande, D., Lakay, F., Shepherd, D. N., van der Walt, E., Lefevre, P., Lett, J.-M., Varsani, A., Rybicki, E. P. & Martin, D. P. (2012).** Adaptive evolution by recombination is not associated with increased mutation rates in Maize streak virus. *BMC evolutionary biology* **12**, 252.
- Morris-Krsinich, B. A. M., Mullineaux, P. M., Donson, J., Boulton, M. I., Markham, P. G., Short, M. N. & Davies, J. W. (1985).** Bidirectional transcription of maize streak virus DNA and identification of the coat protein gene. *Nucleic Acids Research* **13**, 7237-7256.
- Morris, B. A. M., Richardson, K. A., Haley, A., Zhan, X. & Thomas, J. E. (1992).** The nucleotide sequence of the infectious cloned dna component of tobacco yellow dwarf virus reveals features of geminiviruses infecting monocotyledonous plants. *Virology* **187**, 633-642.
- Muhire, B., Golden, M., Murrell, B., Lefevre, P., Lett, J., Gray, A., Poon, A., Kwanele Ngandu, N., Semegni, J. & Tanov, E. (2014).** Evidence of pervasive biologically functional secondary structures within the genomes of eukaryotic single-stranded DNA viruses. *J Gen Virol* **88**, 1972-1989.
- Muhire, B., Martin, D. P., Brown, J. K., Navas-Castillo, J., Moriones, E., Zerbini, M. F., Rivera-Bustamante, R. F., Malathi, V. G., Briddon, R. W. & Varsani, A. (2013).** A

- genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). *Arch Virol* **158**, 1411-1424.
- Mullineaux, P., Donson, J., Morris-Krsinich, B., Boulton, M. & Davies, J. (1984).** The nucleotide sequence of maize streak virus DNA. *The EMBO journal* **3**, 3063.
- Mullineaux, P. M., Guerineau, F. & Accotto, G.-P. (1990).** Processing of complementary sense RNAs of Digitaria streak virus in its host and in transgenic tobacco. *Nucleic Acids Research* **18**, 7259-7265.
- Mumtaz, H., Kumari, S. G., Mansoor, S., Martin, D. P. & Briddon, R. W. (2011).** Analysis of the sequence of a dicot-infecting mastrevirus (family *Geminiviridae*) originating from Syria. *Virus Genes* **42**, 422-428.
- Murad, L., Bielawski, J., Matyasek, R., Kovarik, A., Nichols, R., Leitch, A. & Lichtenstein, C. (2004).** The origin and evolution of geminivirus-related DNA sequences in Nicotiana. *Heredity* **92**, 352-358.
- Nahid, N., Amin, I., Mansoor, S., Rybicki, E., van der Walt, E. & Briddon, R. (2008).** Two dicot-infecting mastreviruses (family *Geminiviridae*) occur in Pakistan. *Arch Virol* **153**, 1441-1451.
- Nash, T. E., Dallas, M. B., Reyes, M. I., Buhrman, G. K., Ascencio-Ibañez, J. T. & Hanley-Bowdoin, L. (2011).** Functional analysis of a novel motif conserved across geminivirus Rep proteins. *Journal of Virology* **85**, 1182-1192.
- Nawaz-ul-Rehman, M. S., Nahid, N., Mansoor, S., Briddon, R. W. & Fauquet, C. M. (2010).** Post-transcriptional gene silencing suppressor activity of two non-pathogenic alphasatellites associated with a begomovirus. *Virology* **405**, 300-308.
- Nelson, J. R., Cai, Y. C., Giesler, T. L., Farchaus, J. W., Sundaram, S. T., Ortiz-Rivera, M., Hosta, L. P., Hewitt, P. L., Mamone, J. A. & Palaniappan, C. (2002).** TempliPhi, phi29 DNA polymerase based rolling circle amplification of templates for DNA sequencing. *BioTechniques* **32**, S44-S47.
- Ng, T. F. F., Duffy, S., Polston, J. E., Bixby, E., Vallad, G. E. & Breitbart, M. (2011a).** Exploring the Diversity of Plant DNA Viruses and Their Satellites Using Vector-Enabled Metagenomics on Whiteflies. *PLoS ONE* **6**, e19050.
- Ng, T. F. F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., Oderinde, B. S., Wommack, K. E. & Delwart, E. (2012).** High Variety of Known and New RNA and DNA Viruses of Diverse Origins in Untreated Sewage. *Journal of Virology* **86**, 12161-12175.
- Ng, T. F. F., Willner, D. L., Lim, Y. W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y., Rohwer, F. & Breitbart, M. (2011b).** Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PLoS ONE* **6**, e20579.
- Niel, C., Diniz-Mendes, L. & Devalle, S. (2005).** Rolling-circle amplification of Torque teno virus (TTV) complete genomes from human and swine sera and identification of a novel swine TTV genogroup. *Journal of General Virology* **86**, 1343-1347.
- Nishigawa, H., Miyata, S.-i., Oshima, K., Sawayanagi, T., Komoto, A., Kuboyama, T., Matsuda, I., Tsuchizaki, T. & Namba, S. (2001).** In planta expression of a protein encoded by the extrachromosomal DNA of a phytoplasma and related to geminivirus replication proteins. *Microbiology* **147**, 507-513.
- Nishigawa, H., Oshima, K., Kakizawa, S., Jung, H.-y., Kuboyama, T., Miyata, S.-i., Ugaki, M. & Namba, S. (2002).** Evidence of intermolecular recombination between

- extrachromosomal DNAs in phytoplasma: a trigger for the biological diversity of phytoplasma? *Microbiology* **148**, 1389-1396.
- Noueiry, A. O., Lucas, W. J. & Gilbertson, R. L. (1994).** Two proteins of a plant DNA virus coordinate nuclear and plasmodesmal transport. *Cell* **76**, 925-932.
- Okoth, V., Dabrowski, Z., Thottappilly, G. & Van Emden, H. (1987).** Comparative analysis of some parameters affecting maize streak virus (MSV) transmission of various *Cicadulina* spp. populations. *International Journal of Tropical Insect Science* **8**, 295-300.
- Oluwafemi, S., Alegbejo, M. D., Onasanya, A. & Olufemi, O. (2011).** Relatedness of Maize streak virus in maize (*Zea mays* L.) to some grass isolates collected from different regions in Nigeria. *African Journal of Agricultural Research* **6**, 5878-5883.
- Oluwafemi, S., Kraberger, S., Shepherd, D. N., Martin, D. P. & Varsani, A. (2014).** A high degree of African streak virus diversity within Nigerian maize fields includes a new mastrevirus from *Axonopus compressus*. *Arch Virol* **159**, 2765-2770.
- Oluwafemi, S., Varsani, A., Monjane, A. L., Shepherd, D. N., Owor, B. E., Rybicki, E. P. & Martin, D. P. (2008).** A new African streak virus species from Nigeria. *Arch Virol* **153**, 1407-1410.
- Orozco, B. M., Gladfelter, H. J., Settlege, S. B., Eagle, P. A., Gentry, R. N. & Hanley-Bowdoin, L. (1998).** Multiple *Cis* Elements Contribute to Geminivirus Origin Function. *Virology* **242**, 346-356.
- Orozco, B. M. & Hanley-Bowdoin, L. (1996).** A DNA structure is required for geminivirus replication origin function. *Journal of Virology* **70**, 148-158.
- Orozco, B. M. & Hanley-Bowdoin, L. (1998).** Conserved Sequence and Structural Motifs Contribute to the DNA Binding and Cleavage Activities of a Geminivirus Replication Protein. *Journal of Biological Chemistry* **273**, 24448-24456.
- Ortmann, A. C. & Suttle, C. A. (2005).** High abundances of viruses in a deep-sea hydrothermal vent system indicates viral mediated microbial mortality. *Deep Sea Research Part I: Oceanographic Research Papers* **52**, 1515-1527.
- Oshima, K., Kakizawa, S., Nishigawa, H., Kuboyama, T., Miyata, S.-i., Ugaki, M. & Namba, S. (2001).** A Plasmid of Phytoplasma Encodes a Unique Replication Protein Having Both Plasmid- and Virus-like Domains: Clue to Viral Ancestry or Result of Virus/Plasmid Recombination? *Virology* **285**, 270-277.
- Owor, B. E., Martin, D. P., Shepherd, D. N., Edema, R., Monjane, A. L., Rybicki, E. P., Thomson, J. A. & Varsani, A. (2007).** Genetic analysis of maize streak virus isolates from Uganda reveals widespread distribution of a recombinant variant. *Journal of General Virology* **88**, 3154-3165.
- Padidam, M., Sawyer, S. & Fauquet, C. M. (1999).** Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218-225.
- Pande, D., Kraberger, S., Lefeuvre, P., Lett, J.-M., Shepherd, D., Varsani, A. & Martin, D. (2012).** A novel maize-infecting mastrevirus from La Réunion Island. *Arch Virol* **157**, 1617-1621.
- Phan, T., Luchsinger, V., Avendano, L., Deng, X. & Delwart, E. (2014).** Cyclovirus in nasopharyngeal aspirates of Chilean children with respiratory infections. *Journal of General Virology* **95**, 922-927.
- Pilartz, M. & Jeske, H. (1992).** Abutilon mosaic geminivirus double-stranded DNA is packed into minichromosomes. *Virology* **189**, 800-802.

- Pita, J., Fondong, V., Sangare, A., Otim-Nape, G., Ogwal, S. & Fauquet, C. (2001).** Recombination, pseudorecombination and synergism of geminiviruses are determinant keys to the epidemic of severe cassava mosaic disease in Uganda. *Journal of General Virology* **82**, 655-665.
- Poojari, S., Alabi, O. J., Fofanov, V. Y. & Naidu, R. A. (2013).** A leafhopper-transmissible DNA virus with novel evolutionary lineage in the family Geminiviridae implicated in grapevine redleaf disease by next-Generation sequencing. *PloS one* **8**, e64194.
- Preiss, W. & Jeske, H. (2003).** Multitasking in replication is common among geminiviruses. *Journal of virology* **77**, 2972-2980.
- Prigent, M., Leroy, M., Confalonieri, F., Dutertre, M. & DuBow, M. S. (2005).** A diversity of bacteriophage forms and genomes can be isolated from the surface sands of the Sahara Desert. *Extremophiles* **9**, 289-296.
- Qin, S., Ward, B. M. & Lazarowitz, S. G. (1998).** The Bipartite Geminivirus Coat Protein Aids BR1 Function in Viral Movement by Affecting the Accumulation of Viral Single-Stranded DNA. *Journal of Virology* **72**, 9247-9256.
- Razavinejad, S. & Heydarnejad, J. (2013).** Transmission and natural hosts of turnip curly top virus. *Iranian Journal of Plant Pathology* **49**, 27-28.
- Razavinejad, S., Heydarnejad, J., Kamali, M., Massumi, H., Kraberger, S. & Varsani, A. (2013).** Genetic diversity and host range studies of turnip curly top virus. *Virus Genes* **46**, 345-353.
- Rekab, D., Carraro, L., Schneider, B., Seemüller, E., Chen, J., Chang, C.-J., Locci, R. & Firrao, G. (1999).** Geminivirus-related extrachromosomal DNAs of the X-clade phytoplasmas share high sequence similarity. *Microbiology* **145**, 1453-1459.
- Ribeiro, S., Ambrozevicius, L., Avila, A. d., Bezerra, I., Calegario, R., Fernandes, J., Lima, M., De Mello, R., Rocha, H. & Zerbini, F. (2003).** Distribution and genetic diversity of tomato-infecting begomoviruses in Brazil. *Arch Virol* **148**, 281-295.
- Rigden, J., Dry, I., Krake, L. & Rezaian, M. (1996).** Plant virus DNA replication processes in Agrobacterium: insight into the origins of geminiviruses? *Proceedings of the National Academy of Sciences* **93**, 10280-10284.
- Roine, E., Kukkaro, P., Paulin, L., Laurinavičius, S., Domanska, A., Somerharju, P. & Bamford, D. H. (2010).** New, closely related haloarchaeal viral elements with different nucleic acid types. *Journal of virology* **84**, 3682-3689.
- Rojas, M. R., Hagen, C., Lucas, W. J. & Gilbertson, R. L. (2005).** Exploiting chinks in the plant's armor: evolution and emergence of geminiviruses. *Annu Rev Phytopathol* **43**, 361-394.
- Rojas, M. R., Noueiry, A. O., Lucas, W. J. & Gilbertson, R. L. (1998).** Bean dwarf mosaic geminivirus movement proteins recognize DNA in a form-and size-specific manner. *Cell* **95**, 105-113.
- Roossinck, M. J. (1997).** Mechanisms of plant virus evolution. *Annual review of phytopathology* **35**, 191-209.
- Rosario, K., Dayaram, A., Marinov, M., Ware, J., Kraberger, S., Stainton, D., Breitbart, M. & Varsani, A. (2012a).** Diverse circular single-stranded DNA viruses discovered in dragonflies (Odonata: Epiprocta). *Journal of General Virology* **93**, 2668-2681.
- Rosario, K., Duffy, S. & Breitbart, M. (2012b).** A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol* **157**, 1851-1871.

- Rosario, K., Marinov, M., Stainton, D., Kraberger, S., Wiltshire, E. J., Collings, D. A., Walters, M., Martin, D. P., Breitbart, M. & Varsani, A. (2011). Dragonfly cyclovirus, a novel single-stranded DNA virus discovered in dragonflies (Odonata: Anisoptera). *Journal of General Virology* **92**, 1302-1308.
- Rosario, K., Nilsson, C., Lim, Y. W., Ruan, Y. & Breitbart, M. (2009a). Metagenomic analysis of viruses in reclaimed water. *Environmental Microbiology* **11**, 2806-2820.
- Rosario, K., Padilla-Rodriguez, M., Kraberger, S., Stainton, D., Martin, D. P., Breitbart, M. & Varsani, A. (2013). Discovery of a novel mastrevirus and alphasatellite-like circular DNA in dragonflies (Ephemeroptera) from Puerto Rico. *Virus Research* **171**, 231-237.
- Rosario, K., Symonds, E. M., Sinigalliano, C., Stewart, J. & Breitbart, M. (2009b). Pepper Mild Mottle Virus as an Indicator of Fecal Pollution. *Applied and Environmental Microbiology* **75**, 7261-7267.
- Rose, D. (1962). Insect vectors of maize streak. *Zool Soc S Afr News Bull* **3**.
- Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime- Ngando, T. & Debroas, D. (2012). Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. *PLoS ONE* **7**, e33641.
- Rybicki, E. (1994). A phylogenetic and evolutionary justification for three genera of Geminiviridae. *Arch Virol* **139**, 49-77.
- Saccardo, F., Cettul, E., Palmano, S., Noris, E. & Firrao, G. (2011). On the alleged origin of geminiviruses from extrachromosomal DNAs of phytoplasmas. *BMC evolutionary biology* **11**, 185.
- Sanz-Burgos, A. P. & Gutiérrez, C. (1998). Organization of the cis-Acting Element Required for Wheat Dwarf Geminivirus DNA Replication and Visualization of a Rep Protein-DNA Complex. *Virology* **243**, 119-129.
- Sanz, A. I., Fraile, A., García-Arenal, F., Zhou, X., Robinson, D. J., Khalid, S., Butt, T. & Harrison, B. D. (2000). Multiple infection, recombination and genome relationships among begomovirus isolates found in cotton and other plants in Pakistan. *Journal of General Virology* **81**, 1839-1849.
- Saunders, K., Lucy, A. & Stanley, J. (1991). DNA forms of the geminivirus African cassava mosaic virus consistent with a rolling circle mechanism of replication. *Nucleic Acids Research* **19**, 2325-2330.
- Saunders, K., Salim, N., Mali, V. R., Malathi, V. G., Briddon, R., Markham, P. G. & Stanley, J. (2002). Characterisation of Sri Lankan cassava mosaic virus and Indian cassava mosaic virus: evidence for acquisition of a DNA B component by a monopartite begomovirus. *Virology* **293**, 63-74.
- Schalk, H. J., Matzeit, V., Schiller, B., Schell, J. & Gronenborn, B. (1989). Wheat dwarf virus, a geminivirus of graminaceous plants needs splicing for replication. *EMBO Journal* **8**, 359-364.
- Schnippenkoetter, W. H., Martin, D. P., Hughes, F. L., Fyvie, M., Willment, J. A., James, D., von Wechmar, M. B. & Rybicki, E. P. (2001). The relative infectivities and genomic characterisation of three distinct mastreviruses from South Africa. *Arch Virol* **146**, 1075-1088.
- Schubert, J., Habekuß, A., Kazmaier, K. & Jeske, H. (2007). Surveying cereal-infecting geminiviruses in Germany—Diagnostics and direct sequencing using rolling circle amplification. *Virus Research* **127**, 61-70.

- Schubert, J., Habekuß, A., Wu, B., Thieme, T. & Wang, X. (2013). Analysis of complete genomes of isolates of the Wheat dwarf virus from new geographical locations and descriptions of their defective forms. *Virus Genes* **48**, 133-139.
- Schwinghamer, M., Thomas, J., Schilg, M., Parry, J., Dann, E., Moore, K. & Kumari, S. (2010). Mastreviruses in chickpea (*Cicer arietinum*) and other dicotyledonous crops and weeds in Queensland and northern New South Wales, Australia. *Australasian Plant Pathology* **39**, 551-561.
- Seguin, J., Rajeswaran, R., Malpica-López, N., Martin, R. R., Kasschau, K., Dolja, V. V., Otten, P., Farinelli, L. & Pooggin, M. M. (2014). De novo reconstruction of consensus master genomes of plant RNA and DNA viruses from siRNAs. *PloS one* **9**, e88513.
- Selth, L. A., Randles, J. W. & Rezaian, M. A. (2002). *Agrobacterium tumefaciens* supports DNA replication of diverse geminivirus types. *FEBS letters* **516**, 179-182.
- Shcherbakov, V. P., Plugina, L., Shcherbakova, T., Sizova, S. & Kudryashova, E. (2011). On the mutagenicity of homologous recombination and double-strand break repair in bacteriophage. *DNA repair* **10**, 16-23.
- Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology* **26**, 1135-1145.
- Shepherd, D. N., Martin, D. P., Lefeuvre, P., Monjane, A. L., Owor, B. E., Rybicki, E. P. & Varsani, A. (2008a). A protocol for the rapid isolation of full geminivirus genomes from dried plant tissue. *Journal of Virological Methods* **149**, 97-102.
- Shepherd, D. N., Martin, D. P., McGivern, D. R., Boulton, M. I., Thomson, J. A. & Rybicki, E. P. (2005). A three-nucleotide mutation altering the Maize streak virus Rep pRBR-interaction motif reduces symptom severity in maize and partially reverts at high frequency without restoring pRBR–Rep binding. *Journal of general virology* **86**, 803-813.
- Shepherd, D. N., Martin, D. P., Van Der Walt, E., Dent, K., Varsani, A. & Rybicki, E. P. (2010). Maize streak virus: An old and complex 'emerging' pathogen. *Molecular Plant Pathology* **11**, 1-12.
- Shepherd, D. N., Martin, D. P., Varsani, A., Thomson, J. A., Rybicki, E. P. & Klump, H. H. (2006). Restoration of native folding of single-stranded DNA sequences through reverse mutations: An indication of a new epigenetic mechanism. *Archives of Biochemistry and Biophysics* **453**, 108-122.
- Shepherd, D. N., Varsani, A., Windram, O. P., Lefeuvre, P., Monjane, A. L., Owor, B. E. & Martin, D. P. (2008b). Novel sugarcane streak and sugarcane streak Reunion mastreviruses from southern Africa and la Réunion. *Arch Virol* **153**, 605-609.
- Sikorski, A., Argüello-Astorga, G., Dayaram, A., Dobson, R. J. & Varsani, A. (2012). Discovery of a novel circular single-stranded DNA virus from porcine faeces. *Arch Virol* **158**, 283-289.
- Sikorski, A., Massaro, M., Kraberger, S., Young, L. M., Smalley, D., Martin, D. P. & Varsani, A. (2013). Novel myco-like DNA viruses discovered in the faecal matter of various animals. *Virus Research* **177**, 209-216.
- Simons, J. N. & Coe, D. M. (1958). Transmission of pseudo-curly top virus in Florida by a treehopper. *Virology* **6**, 43-48.
- Smits, S. L., Zijlstra, E. E., van Hellemond, J. J., Schapendonk, C., Bodewes, R., Schürch, A. C., Haagmans, B. L. & Osterhaus, A. (2012). Novel cyclovirus in human cerebrospinal fluid, Malawi, 2010-2011. *Emerging Infectious Diseases* **19**, 1511.

- Soffer, N., Brandt, M. E., Correa, A. M., Smith, T. B. & Thurber, R. V. (2013).** Potential role of viruses in white plague coral disease. *The ISME journal* **8**, 271-283.
- Soleimani, R., Matic, S., Taheri, H., Behjatnia, S. A. A., Vecchiati, M., Izadpanah, K. & Accotto, G. P. (2013).** The unconventional geminivirus Beet curly top Iran virus: satisfying Koch's postulates and determining vector and host range. *Annals of Applied Biology* **162**, 174-181.
- Stanley, J. (1995).** Analysis of African cassava mosaic virus recombinants suggests strand nicking occurs within the conserved nonanucleotide motif during the initiation of rolling circle DNA replication. *Virology* **206**, 707-712.
- Stanley, J., Markham, P., Callis, R. & Pinner, M. (1986).** The nucleotide sequence of an infectious clone of the geminivirus beet curly top virus. *The EMBO journal* **5**, 1761.
- Stenger, D. (1993).** Complete nucleotide sequence of the hypervirulent CFH strain of beet curly top virus. *Molecular plant-microbe interactions: MPMI* **7**, 154-157.
- Storey, H. (1924).** The transmission of a new plant virus disease by insects. *Nature* **114**, 245.
- Storey, H. (1925).** The transmission of streak disease of maize by the leafhopper *balclutha mbilalide*. *Annals of Applied Biology* **12**, 422-439.
- Storey, H. (1936).** Virus Diseases of East African Plants: IV. A Survey of the Viruses attacking Gramineae. *East African Agricultural Journal* **1**, 333-337.
- Suárez-López, P., Martínez-Salas, E., Hernández, P. & Gutiérrez, C. (1995).** Bent DNA in the large intergenic region of wheat dwarf geminivirus. *Virology* **208**, 303-311.
- Sudarshana, M., Wang, H., Lucas, W. & Gilbertson, R. (1998).** Dynamics of bean dwarf mosaic geminivirus cell-to-cell and long-distance movement in *Phaseolus vulgaris* revealed, using the green fluorescent protein. *Molecular plant-microbe interactions* **11**, 277-291.
- Swanson, M. M., Reavy, B., Makarova, K. S., Cock, P. J., Hopkins, D. W., Torrance, L., Koonin, E. V. & Taliansky, M. (2012).** Novel bacteriophages containing a genome of another bacteriophage within their genomes. *PloS one* **7**, e40683.
- Takacs, C. & Priscu, J. (1998).** Bacterioplankton dynamics in the McMurdo Dry Valley lakes, Antarctica: production and biomass loss over four seasons. *Microbial Ecology* **36**, 239-250.
- Thomas, J., Parry, J., Schwinghamer, M. & Dann, E. (2010).** Two novel mastreviruses from chickpea (*Cicer arietinum*) in Australia. *Arch Virol* **155**, 1777-1788.
- Tobias, I., Shevchenko, O., Kiss, B., Bysov, A., Snihur, H., Polischuk, V., Salanki, K. & Palkovics, L. (2011).** Comparison of the nucleotide sequences of wheat dwarf virus (WDV) isolates from Hungary and Ukraine. *Polish Journal of Microbiology* **60**, 125-131.
- Tran-Nguyen, L. & Gibb, K. (2006).** Extrachromosomal DNA isolated from tomato big bud and *Candidatus* *Phytoplasma australiense* phytoplasma strains. *Plasmid* **56**, 153-166.
- Trębicki, P., Harding, R., Rodoni, B., Baxter, G. & Powell, K. (2010).** Vectors and alternative hosts of Tobacco yellow dwarf virus in southeastern Australia. *Annals of Applied Biology* **157**, 13-24.
- Unsold, S., Höhnle, M., Ringel, M. & Frischmuth, T. (2001).** Subcellular Targeting of the Coat Protein of African Cassava Mosaic Geminivirus. *Virology* **286**, 373-383.

- van den Brand, J. M. A., van Leeuwen, M., Schapendonk, C. M., Simon, J. H., Haagmans, B. L., Osterhaus, A. D. M. E. & Smits, S. L. (2012). Metagenomic Analysis of the Viral Flora of Pine Marten and European Badger Feces. *Journal of Virology* **86**, 2360-2365.
- Van der Walt, E., Martin, D. P., Varsani, A., Polston, J. E. & Rybicki, E. P. (2008a). Experimental observations of rapid Maize streak virus evolution reveal a strand-specific nucleotide substitution bias. *Virology journal* **5**, 104.
- van der Walt, E., Palmer, K. E., Martin, D. P. & Rybicki, E. P. (2008b). Viable chimaeric viruses confirm the biological importance of sequence specific maize streak virus movement protein and coat protein interactions. *Virol J* **5**, 61.
- van der Walt, E., Rybicki, E. P., Varsani, A., Polston, J. E., Billharz, R., Donaldson, L., Monjane, A. L. & Martin, D. P. (2009). Rapid host adaptation by extensive recombination. *Journal of General Virology* **90**, 734-746.
- van Doorn, H. R., Nghia, H. D. T., Chau, T. T. H., de Vries, M., Canuti, M., Deijs, M., Jebbink, M. F., Baker, S., Bryant, J. E. & Tham, N. T. (2013). Identification of a new cyclovirus in cerebrospinal fluid of patients with acute central nervous system infections. *MBio* **4**, e00231-00213.
- Varma, A. & Malathi, V. (2003). Emerging geminivirus problems: a serious threat to crop production. *Annals of Applied Biology* **142**, 145-164.
- Varsani, A., Martin, D. P., Navas-Castillo, J., Moriones, E., Hernández-Zepeda, C., Idris, A., Zerbini, F. M. & Brown, J. K. (2014a). Revisiting the classification of curtoviruses based on genome-wide pairwise identity. *Arch Virol* **159**, 1873-1882.
- Varsani, A., Monjane, A. L., Donaldson, L., Oluwafemi, S., Zinga, I., Komba, E. K., Plakoutene, D., Mandakombo, N., Mboukoulida, J., Semballa, S., Briddon, R. W., Markham, P. G., Lett, J. M., Lefeuvre, P., Rybicki, E. P. & Martin, D. P. (2009a). Comparative analysis of Panicum streak virus and Maize streak virus diversity, recombination patterns and phylogeography. *Virology Journal* **6**, 194.
- Varsani, A., Navas-Castillo, J., Moriones, E., Hernández-Zepeda, C., Idris, A., Brown, J., Murilo Zerbini, F. & Martin, D. (2014b). Establishment of three new genera in the family Geminiviridae: Becurtovirus, Eragrovirus and Turncurovirus. *Arch Virol* **159**, 2193-2203.
- Varsani, A., Oluwafemi, S., Windram, O., Shepherd, D., Monjane, A., Owor, B., Rybicki, E., Lefeuvre, P. & Martin, D. (2008a). Panicum streak virus diversity is similar to that observed for maize streak virus. *Arch Virol* **153**, 601-604.
- Varsani, A., Shepherd, D. N., Dent, K., Monjane, A. L., Rybicki, E. P. & Martin, D. P. (2009b). A highly divergent South African geminivirus species illuminates the ancient evolutionary history of this family. *Virology Journal* **6**.
- Varsani, A., Shepherd, D. N., Monjane, A. L., Owor, B. E., Erdmann, J. B., Rybicki, E. P., Peterschmitt, M., Briddon, R. W., Markham, P. G., Oluwafemi, S., Windram, O. P., Lefeuvre, P., Lett, J. M. & Martin, D. P. (2008b). Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. *Journal of General Virology* **89**, 2063-2074.
- Velásquez-Valle, R., Mena-Covarrubias, J., Reveles-Torres, L., Argüello-Astorga, G., Salas-Luevano, M. & Mauricio-Castillo, J. (2012). First report of Beet mild curly top virus in dry bean in Zacatecas, Mexico. *Plant Disease* **96**, 771-771.

- Victoria, J. G., Kapoor, A., Li, L., Blinkova, O., Slikas, B., Wang, C., Naeem, A., Zaidi, S. & Delwart, E. (2009).** Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *Journal of Virology* **83**, 4642-4651.
- Wang, Y., Dang, M., Hou, H., Mei, Y., Qian, Y. & Zhou, X. (2014a).** Identification of an RNA Silencing Suppressor encoded by a Mastrevirus. *Journal of General Virology* **95**, 2082-2088.
- Wang, Y., Mao, Q., Liu, W., Mar, T. T., Wei, T., Liu, Y. & Wang, X. (2014b).** Localization and distribution of Wheat dwarf virus in its vector leafhopper, *Psammotettix alienus*. *Phytopathology* **104**, 897-904.
- Webb, M. (1987).** Species recognition in Cicadulina leafhoppers (Hemiptera: Cicadellidae), vectors of pathogens of Gramineae. *Bulletin of entomological research* **77**, 683-712.
- Whon, T. W., Kim, M.-S., Roh, S. W., Shin, N.-R., Lee, H.-W. & Bae, J.-W. (2012).** Metagenomic Characterization of Airborne Viral DNA Diversity in the Near-Surface Atmosphere. *Journal of Virology* **86**, 8221-8231.
- Williamson, S. J., Cary, S. C., Williamson, K. E., Helton, R. R., Bench, S. R., Winget, D. & Wommack, K. E. (2008).** Lysogenic virus–host interactions predominate at deep-sea diffuse-flow hydrothermal vents. *The ISME journal* **2**, 1112-1121.
- Willment, J. A., Martin, D. P., Palmer, K. E., Schnippenkoetter, W. H., Shepherd, D. N. & Rybicki, E. P. (2007).** Identification of long intergenic region sequences involved in maize streak virus replication. *Journal of General Virology* **88**, 1831-1841.
- Willment, J. A., Martin, D. P. & Rybicki, E. P. (2001).** Analysis of the diversity of African streak mastreviruses using PCR-generated RFLPs and partial sequence data. *Journal of Virological Methods* **93**, 75-87.
- Wright, E. A., Heckel, T., Groenendijk, J., Davies, J. W. & Boulton, M. I. (1997).** Splicing features in maize streak virus virion- and complementary-sense gene expression. *The Plant Journal* **12**, 1285-1297.
- Wu, B., Melcher, U., Guo, X., Wang, X., Fan, L. & Zhou, G. (2008).** Assessment of codivergence of Mastreviruses with their plant hosts. *BMC evolutionary biology* **8**, 335.
- Xie, O., Suarez-Lopez, P. & Gutierrez, C. (1995).** Identification and analysis of a retinoblastoma binding motif in the replication protein of a plant DNA virus: Requirement for efficient viral DNA replication. *EMBO Journal* **14**, 4073-4082.
- Xie, Q., Sanz-Burgos, A. P., Guo, H., García, J. A. & Gutiérrez, C. (1999).** GRAB proteins, novel members of the NAC domain family, isolated by their interaction with a geminivirus protein. *Plant molecular biology* **39**, 647-656.
- Yazdi, H., Heydarnejad, J. & Massumi, H. (2008).** Genome characterization and genetic diversity of beet curly top Iran virus: a geminivirus with a novel nonanucleotide. *Virus Genes* **36**, 539-545.
- Yoshida, M., Takaki, Y., Eitoku, M., Nunoura, T. & Takai, K. (2013).** Metagenomic Analysis of Viral Communities in (Hado)Pelagic Sediments. *PLoS ONE* **8**, e57271.
- Yu, X., Li, B., Fu, Y., Jiang, D., Ghabrial, S. A., Li, G., Peng, Y., Xie, J., Cheng, J., Huang, J. & Yi, X. (2010).** A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. *Proceedings of the National Academy of Sciences* **107**, 8387-8392.
- Yu, X., Li, B., Fu, Y., Xie, J., Cheng, J., Ghabrial, S. A., Li, G., Yi, X. & Jiang, D. (2013).** Extracellular transmission of a DNA mycovirus and its use as a natural fungicide. *Proceedings of the National Academy of Sciences* **110**, 1452-1457.

- Zawar-Reza, P., Argüello-Astorga, G. R., Krabberger, S., Julian, L., Stainton, D., Broady, P. A. & Varsani, A. (2014).** Diverse small circular single-stranded DNA viruses identified in a freshwater pond on the McMurdo Ice Shelf (Antarctica). *Infection, Genetics and Evolution* **26**, 132-138.
- Zhan, X., Richardson, K. A., Haley, A. & Morris, B. A. (1993).** The activity of the coat protein promoter of chloris striate mosaic virus is enhanced by its own and C1-C2 gene products. *Virology* **193**, 498-502.
- Zhang, J., Chiodini, R., Badr, A. & Zhang, G. (2011).** The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics* **38**, 95-109.
- Zhang, W., Olson, N. H., Baker, T. S., Faulkner, L., Agbandje-McKenna, M., Boulton, M. I., Davies, J. W. & McKenna, R. (2001).** Structure of the Maize Streak Virus Geminata Particle. *Virology* **279**, 471-477.
- Zhou, X. (2013).** Advances in understanding begomovirus satellites. *Annual review of phytopathology* **51**, 357-381.
- Zhou, X., Liu, Y., Calvert, L., Munoz, C., Otim-Nape, G. W., Robinson, D. J. & Harrison, B. D. (1997).** Evidence that DNA-A of a geminivirus associated with severe cassava mosaic disease in Uganda has arisen by interspecific recombination. *Journal of General Virology* **78**, 2101-2111.
- Zhou, X., Liu, Y., Robinson, D. J. & Harrison, B. D. (1998).** Four DNA-A variants among Pakistani isolates of cotton leaf curl virus and their affinities to DNA-A of geminivirus isolates from okra. *Journal of General Virology* **79**, 915-923.

Chapter 2

Australian monocot-infecting mastrevirus diversity rivals that in Africa

Contents

2.1	Abstract.....	76
2.2	Introduction.....	77
2.3	Materials and methods.....	79
2.3.1	Sample collection and virus amplification	79
2.3.2	Sequence and phylogenetic analyses.....	80
2.3.3	Recombination analysis.....	80
2.3.4	Selection analysis	81
2.4	Results and discussion	83
2.4.1	Genome organisation and conserved motifs.....	83
2.4.2	Classification of novel Australian monocot-infecting mastreviruses.....	92
2.4.3	SSMV-1 and SSMV-2 resemble divergent African streak viruses	99
2.4.4	Evidence of inter- and intra- species recombination	102
2.4.5	Selection analyses.....	104
2.4.6	Host range analysis.....	105
2.5	Concluding remarks.....	107

This body of work has been published in Virus Research and is presented in a similar manner to that of the publication:

Kraberger, S., Thomas, J.E., Geering, A.D.W., Dayaram, A., Stainton, D., Hadfield, J., Walters, M., Parmenter, K.S., van Brunschot, S., Collings, D.A., Martin, D.P. and Varsani, A. (2012) Australian monocot-infecting mastrevirus diversity rivals that in Africa. Virus Research 169(1), 127-136.

2.1 Abstract

Monocotyledonous plant-infecting mastreviruses (family *Geminiviridae*) are found in the Old World. The greatest diversity of these viruses to date has been documented in Africa, however this may simply reflect the more extensive sampling that has been done there. To provide a better understanding of mastrevirus diversity in Australia, we have sequenced the genomes of 41 virus isolates found in naturalised and native grasses and identified four tentative new species in addition to the four previously characterised species. Two of these species, which we have tentatively named *Sporobolus striate mosaic virus 1* and *2* (SSMV-1 and SSMV-2), were recovered from a single *Sporobolus* plant. Both species are highly divergent and are most closely related to the African streak viruses. This information, coupled with the discovery of divergent dicotyledonous plant infecting mastreviruses in Australia brings into question the hypothesis that mastreviruses may have originated in Africa. We found that the patterns of inter- and intra-species recombination and the recombination hotspots mirror those found in both their African monocot-infecting counterparts and dicot-infecting mastrevirus.

2.2 Introduction

Viruses in the family *Geminiviridae* cause diseases in a wide range of economically important domesticated plant species. The four currently described genera, *Mastrevirus*, *Curtovirus*, *Begomovirus* and *Topocuvirus*, differ in genome organisation, host range and insect vector species. A few currently unclassified and divergent geminivirus species also exist, all with unknown vector species (Brown *et al.*, 2011). These include *Beet curly top Iran virus* (BCTIV) (Yazdi *et al.*, 2008) Eragrostis curvula streak virus (ECSV) (Varsani *et al.*, 2009b) and Turnip curly top virus (TCTV) (Briddon *et al.*, 2010a) and these may in the future be classified as members of three new geminivirus genera. Geminiviruses characteristically have a small ~2.6–2.8 kb circular, single-stranded DNA genome (bipartite begomoviruses have two DNA molecules of ~2.6–2.8 kb each), encapsidated in twinned isometric viral particles (Harrison, 1985). The genomes are bidirectionally transcribed and encode between four (*Mastrevirus* and ECSV) and eight (*Curtovirus*, *Begomovirus* and *Topocuvirus*, BCTIV and TCTV) genes, and replicate in the cell nucleus via rolling circle and recombination-dependent mechanisms (Jeske *et al.*, 2001; Saunders *et al.*, 1991; Stenger *et al.*, 1991). Although not the most populated geminivirus genus, the genus *Mastrevirus* contains a wide diversity of viruses, are transmitted by a wider variety of vector species (including at least 14 different leafhopper species spread across at least five genera) and infect a greater diversity of host species than any of the other geminivirus genera, with their hosts including both monocot and dicot angiosperms.

All known mastreviruses have a monopartite 2.5–2.7 kb genome encoding four genes. These are movement protein (*mp*) and a coat protein (*cp*) on the virion-sense strand and, on the complementary strand, a replication associated protein (*rep*) and a replication associated protein A (*repA*) genes. The *rep* and *repA* are expressed from spliced complementary strand transcripts (Dekker *et al.*, 1991; Mullineaux *et al.*, 1990; Schalk *et al.*, 1989; Wright *et al.*, 1997). As with all other geminiviruses, the Rep initiates rolling circle replication by binding close to and then nicking the origin of replication at a hairpin structure between nucleotides (nts) 7 and 8 of a conserved nonanucleotide (TAAT[A/G]TTAC) loop sequence (Heyraud *et al.*, 1993).

To date, fifteen species of monocot-infecting mastreviruses have been characterised from the Old World (including ten from Africa and nearby south-west Indian Ocean islands, four from Eurasia, four from Australia and one from Vanuatu) but none from the New World. Dicot-infecting mastreviruses have also been found in the Middle East (Mumtaz *et al.*, 2011), Pakistan (Nahid *et al.*, 2008), Africa (Liu *et al.*, 1997) and Australia (Hadfield *et al.*, 2012; Schwinghamer *et al.*, 2010; Thomas *et al.*, 2010). Since the only monocot-infecting mastrevirus with any economic importance is Maize streak virus (MSV) from Africa, very little is known about the diversity, distributions and host ranges of monocot-infecting mastreviruses in other regions of the world. In Australia, biological characterizations involving host range analysis (Greber, 1989), serological studies (Pinner *et al.*, 1992) and genome sequencing have identified the existence of four distinct Australian monocot-infecting mastrevirus species. These are *Chloris striate mosaic virus* (CSMV) (Andersen *et al.*, 1988), *Digitaria didactyla striate mosaic virus* (DDSMV) (Briddon *et al.*, 2010b), *Bromus catharticus striate mosaic virus* (BCSMV) (Hadfield *et al.*, 2011) and *Paspalum striate mosaic virus* (PSMV) (Geering *et al.*, 2011).

Just as more intensive sampling and sequencing have begun to reveal the true extent of African monocot-infecting mastrevirus diversity and evolutionary mechanisms (Martin *et al.*, 2011a; Martin *et al.*, 2001; Monjane *et al.*, 2011; Varsani *et al.*, 2008b; Willment *et al.*, 2001), it is likely that similar efforts in Australia will be equally productive. In this study, we have determine the full genome sequences of, and recombination patterns within, forty one Australian monocot-infecting mastrevirus isolates from 40 symptomatic wild grass samples collected from various locations in Queensland and New South Wales and including four novel species.

2.3 Materials and methods

2.3.1 Sample collection and virus amplification

Grass samples showing either streak or striation mosaic-like patterns (n=40) were collected from various locations around Queensland and New South Wales in eastern Australia between 1983 and 2011. Of these samples, 39 were collected between 2003 and 2011, while the other two samples were collected in 1983 and 1998 (Table 2.1).

Total DNA was extracted from dried plant material using the GenCatch Plant Genomic DNA Purification kit (Epoch Biolabs, USA). Circular viral DNA in the total DNA extract was amplified using the Illustra TempliPhi Amplification Kit (GE Healthcare, USA) as described by Owor *et al.* (2007b) and (Shepherd *et al.*, 2008a). The amplified concatenated viral DNA was subsequently digested using the restriction enzymes *Pst*I and *Bgl*II to yield unit length genomes. The resulting ~2.7 kb DNA fragments were purified using the MEGA-spin Agarose Gel Extraction Kit (iNtRON Biotechnology Inc., Korea) and ligated into *Pst*I and *Bgl*II sites of the pGEM3Zf(+) vector (Promega Biotech, USA).

In thirty cases, polymerase chain reaction (PCR) allowed the recovery of viral genomes from samples where no unique restriction sites could be determined. For these samples, we designed back-to-back primers as follows: forward 5'-GANTTGGTCCGCAGTGTAGA-3', dicot reverse 5'-GTACCGGWAAGACMWCYTGG-3'. The PCR was performed using Kapa HiFi HotStart DNA polymerase (Kapa Biosystems, USA) and using following the thermocycling conditions: 94°C for 3 min, 25 cycles of 98°C (3 min), 52°C (30 sec), 72°C (2.45 min) and a final extension of 72°C for 3 min. PCR products were cloned using the pJET1.2 vector (CloneJET™ PCR cloning kit, Fermentas, USA). In order to identify the host species a section of the chloroplast *ndhF* gene (~1.1kb) was PCR amplified from total genomic DNA using the following primer pairs *ndhF1F*: forward 5'-ATG GAA CAK ACA TAT SAA TAT G-3', *ndhF972R*: reverse 5'-CAT CAT ATA AAC CCA ATT GAG AC-3' and *ndhF972F*: forward 5'-GT CTC AAT TGG GTT ATA TGA T-3', *ndhF2110R*: reverse 5'-CCC CCT AYA TAT TTG ATA CCT T-3'. The PCR was performed using Kapa HiFi HotStart DNA polymerase (Kapa Biosystems, USA) and the following PCR protocol: 94°C for 3 min, 25 cycles of 98°C (3 min), 48°C (30 sec), 72°C (1.30 min) and a final extension of

72°C for 3 min. PCR products were cloned using the pJET1.2 vector (CloneJET™ PCR cloning kit, Fermentas, USA) and sequenced. All clones were sequenced by Macrogen Inc. (Korea) by primer walking.

2.3.2 Sequence and phylogenetic analyses

Contigs were assembled and general editing done using DNAMAN (version 5.2.9; Lynnon Biosoft) and MEGA5 (Tamura *et al.*, 2011). Open reading frames (ORFs) were identified using DNAMAN, and conceptual translations done using MEGA5. Representatives of previously identified mastrevirus species (both monocot- and dicot-infecting) from throughout the world were obtained from the NCBI GenBank database and used in the full genome and protein (Rep and CP) analyses.

All mastrevirus full genome sequences were linearised at the origin of replication (TAAT(A/G)TTAC) and aligned using MUSCLE (Edgar, 2004). Maximum likelihood (ML) phylogenetic trees were constructed using PHYML version 3 (Guindon *et al.*, 2010) with 1000 non-parametric bootstrap replicates and the nucleotide substitution model GTR+G4 (selected as the best fit model by RDP4) (Martin *et al.*, 2010). ML phylogenetic trees of Rep and CP amino acid sequences were constructed with PHYML using the LG model (previously determined as a best fit model) (Le & Gascuel, 2008). The ML phylogenetic trees were visualised using MEGA5. Branches with <60% support were collapsed using Mesquite (Version 2.75). The phylogenetic trees were rooted with the BCTIV Rep and CP sequences. Pairwise similarity comparisons of full genomes (nucleotide sequences), Rep (amino acid sequences) and CP (amino acid sequences) were undertaken with MEGA 5 (p-distance with pairwise deletion of gaps).

2.3.3 Recombination analysis

Recombination within the Australian monocot-infecting mastreviruses was analysed using the following methods in the RDP4 software package (Martin *et al.*, 2010): RDP (Martin & Rybicki, 2000), GENECONV (Padidam *et al.*, 1999), Bootscan (Martin *et al.*, 2005), Maxchi (Smith, 1992), Chimera (Posada & Crandall, 2001), Siscan (Gibbs *et al.*, 2000), and 3Seq (Boni *et al.*, 2007). Datasets were grouped appropriately to scan for inter- and intra- species recombination. Recombination events were considered to be authentic when detected by a minimum of three methods coupled with clear phylogenetic evidence.

2.3.4 Selection analysis

The ratios of normalised synonymous (dS) and non-synonymous (dN) substitution rates from codon alignments of *mp*, *cp* and *rep* gene of CSMV, PSMV and combined dataset of PSMV/PDSMV/DCSMV/BCSMV were calculated taking recombination into account using the SLAC method implemented in the online server Datamonkey (<http://www.datamonkey.org>).

Table 2.1: Details of Australian monocot-infecting mastrevirus isolates from this study (41 isolates) and previous studies (4 isolates). QLD-Queensland and NSW-New South Wales.

	Isolate	Sampling year	GenBank accession	Host	Location	Latitude	Longitude
Current study							
1	CSMV-A ₁ [AU-2935-2011]	2011	JQ948053	<i>Chloris gayana</i>	Tolga, QLD	-17.2233	145.4803
2	CSMV-A ₁ [AU-3017-2011]	2011	JQ948054	<i>Eriochloa polystachya</i>	Wee Waa, NSW	-30.2239	149.4465
3	CSMV-A ₂ [AU-1610-2003]	2003	JQ948056	<i>Chloris gayana</i>	Brisbane, QLD	-27.4698	153.0213
4	CSMV-A ₂ [AU-1658-2004]	2004	JQ948057	<i>Paspalum dilatatum</i>	Peak Crossing, QLD	-27.7839	152.7284
5	CSMV-A ₂ [AU-KP11-1983]	1983	JQ948055	<i>Triticum aestivum</i>	Brisbane, QLD	-27.4698	153.0213
6	CSMV-A ₃ [AU-1650-2004]	2004	JQ948058	<i>Chloris gayana</i>	Mt. Glorious, QLD	-27.3343	152.7678
7	CSMV-A ₄ [AU-1649-2004]	2004	JQ948059	<i>Chloris gayana</i>	Beaudesert, QLD	-27.9876	152.9954
8	CSMV-A ₅ [AU-3009-2011]	2011	JQ948060	<i>Chloris gayana</i>	Farrant Hill, NSW	-28.3181	153.4877
9	CSMV-A ₆ [AU-QG29-2011]	2011	JQ948083	<i>Panicum</i> sp.	Wappa Falls, QLD	-26.5738	152.9401
10	CSMV-A ₆ [AU-QG31-2011]	2011	JQ948081	<i>Sporobolus</i> sp.	Wappa Falls, QLD	-26.5738	152.9401
11	CSMV-A ₆ [AU-QG32-2011]	2011	JQ948082	<i>Digitaria ciliaris</i>	Wappa Falls, QLD	-26.5738	152.9401
12	CSMV-A ₆ [AU-QG36-2011]	2011	JQ948084	<i>Chloris gayana</i>	Glasshouse Mountains, QLD	-26.9273	152.9407
13	DCSMV-A [AU-QG6-2011]	2011	JQ948089	<i>Digitaria ciliaris</i>	Corinda, QLD	-27.5509	152.9799
14	DCSMV-A [AU-QG7-2011]	2011	JQ948090	<i>Digitaria ciliaris</i>	Corinda, QLD	-27.5509	152.9799
15	DCSMV-A [AU-QG8-2011]	2011	JQ948091	<i>Digitaria ciliaris</i>	Corinda, QLD	-27.5509	152.9799
16	DCSMV-B [AU-QG5-2011]	2011	JQ948088	<i>Digitaria ciliaris</i>	Corinda, QLD	-27.5509	152.9799
17	PDSMV-A ₁ [AU-QG46-2011]	2011	JQ948087	<i>Paspalum dilatatum</i>	Landsborough, QLD	-26.8079	152.9636
18	PDSMV-A ₂ [AU-QG45-2011]	2011	JQ948086	<i>Paspalum dilatatum</i>	Wappa Dam, QLD	-26.5692	152.9199
19	PDSMV-A ₃ [AU-1652-2004]	2004	JQ948062	<i>Paspalum dilatatum</i>	Brisbane, QLD	-27.4698	153.0213
20	PDSMV-A ₃ [AU-1660-2004]	2004	JQ948061	<i>Paspalum dilatatum</i>	Mt. Glorious, QLD	-27.3343	152.7678
21	PDSMV-A ₃ [AU-QG24-2011]	2011	JQ948077	<i>Paspalum dilatatum</i>	Moorooka, QLD	-27.5313	153.0239
22	PDSMV-A ₃ [AU-QG44-2011]	2011	JQ948085	<i>Digitaria ciliaris</i>	Brisbane Botanical Gardens, QLD	-26.5692	153.0307
23	PSMV-A ₁ [AU-1659-2004]	2004	JQ948063	<i>Paspalum dilatatum</i>	Fernvale, QLD	-27.4560	152.6533
24	PSMV-A ₂ [AU-1656-2004]	2004	JQ948064	<i>Paspalum dilatatum</i>	Beaudesert, QLD	-27.9876	152.9954
25	PSMV-A ₃ [AU-1654-2004]	2004	JQ948065	<i>Paspalum dilatatum</i>	North Maclean, QLD	-27.7721	153.0172
26	PSMV-A ₄ [AU-QG33-2011]	2011	JQ948079	<i>Paspalum dilatatum</i>	Landsborough, QLD	-26.8078	152.9636
27	PSMV-A ₆ [AU-1657-2004]	2004	JQ948068	<i>Paspalum dilatatum</i>	Boonah, QLD	-27.9973	152.6822
28	PSMV-A ₆ [AU-QG12-2011]	2011	JQ948066	<i>Paspalum dilatatum</i>	Corinda, QLD	-27.5509	152.9799
29	PSMV-A ₆ [AU-QG13-2011]	2011	JQ948067	<i>Paspalum dilatatum</i>	Corinda, QLD	-27.5509	152.9799
30	PSMV-A ₆ [AU-QG14-2011]	2011	JQ948071	<i>Paspalum dilatatum</i>	Corinda, QLD	-27.5509	152.9799
31	PSMV-A ₆ [AU-QG15-2011]	2011	JQ948072	<i>Paspalum dilatatum</i>	Corinda, QLD	-27.5509	152.9799
32	PSMV-A ₆ [AU-QG16-2011]	2011	JQ948073	<i>Paspalum dilatatum</i>	Palmgrove National park, QLD	-24.9301	149.4030
33	PSMV-A ₆ [AU-QG17-2011]	2011	JQ948074	<i>Ehrharta erecta</i>	Moorooka, QLD	-27.5313	153.0239
34	PSMV-A ₆ [AU-QG19-2011]	2011	JQ948075	<i>Digitaria ciliaris</i>	Moorooka, QLD	-27.5313	153.0239
35	PSMV-A ₆ [AU-QG2-2011]	2011	JQ948076	<i>Paspalum dilatatum</i>	Corinda, QLD	-27.5509	152.9799
36	PSMV-A ₆ [AU-QG28-2011]	2011	JQ948078	<i>Paspalum dilatatum</i>	Wappa Falls, QLD	-26.5738	152.9401
37	PSMV-A ₆ [AU-QG3-2011]	2011	JQ948080	<i>Paspalum dilatatum</i>	Corinda, QLD	-27.5509	152.9799
38	PSMV-B ₁ [AU-3011-2011]	2011	JQ948069	<i>Paspalum dilatatum</i>	Farrant Hill, NSW	-28.3181	153.4877
39	PSMV-B ₂ [AU-846-1998]	1998	JQ948070	<i>Paspalum dilatatum</i>	Anstead, QLD	-27.5948	152.8449
40	SSMV-1 [AU-3020_1-2011]	2011	JQ948051	<i>Sporobolus</i> sp.	Wyaga, QLD	-28.1826	150.6601
41	SSMV-2 [AU-3020_2-2011]	2011	JQ948052	<i>Sporobolus</i> sp.	Wyaga, QLD	-28.1826	150.6601
Previous studies							
Andersen <i>et al.</i> , 1988	CSMV-A ₂ [AU-QL]	unknown	M20021	<i>Chloris gayana</i>	Brisbane, QLD	-27.3246	152.5175
Geering <i>et al.</i> , 2012	PSMV-A ₅ [AU-1611-2003]	2003	JF905486	<i>Paspalum dilatatum</i>	Anstead, QLD	-27.3246	152.5175
Briddon <i>et al.</i> , 2010	DDSMV [AU-QL-1999]	1999	HM122238	<i>Digitaria didactyla</i>	Brisbane, QLD	-27.4698	153.0213
Hadfield <i>et al.</i> , 2011	BCSMV [AU-QL-1999]	1999	HQ113104	<i>Bromus catharticus</i>	Darling Downs, QLD	-27.5307	150.5852

2.4 Results and discussion

2.4.1 Genome organisation and conserved motifs

The genomes of all 41 of the monocot-infecting mastreviruses were between 2748 and 2818 nucleotides with all containing two intergenic regions (the long intergenic region, LIR, and the short intergenic region, SIR) and open reading frames (ORFs) with homology to the four characteristic mastrevirus genes: on the virion-sense strand, V1 and V2, encoding the *cp* and *mp*, respectively, and on the complementary-sense strand, C1 and C2, encoding the *rep* and the C-terminal part of the *repA*, respectively (Fig. 2.1). We also identified likely intron donor and acceptor sites within C1 which, when spliced, would result in a C1:C2 transcript that could act as the messenger RNA for the *repA* (Accotto *et al.*, 1989). Within all the genomes, we identified the conserved sequence TAATATT↓AC (↓ indicates the site that is likely nicked by the *rep*), located at the putative origin of replication (*v-ori*).

Within all the Australian monocot-infecting mastrevirus *mps*, we identified the putative hydrophobic membrane domain and within the *cp*, the putative DNA binding domain (Fig. 2.2). All the Australian monocot-infecting mastreviruses contained the four rolling circle replication associated motifs: motifs I, II, III and the GRS (see Fig. 2.2 for details). Finally, we identified a conserved LxCxE motif in SSMV-1 and SSMV-2 that likely binds the host's retinoblastoma related protein (pRBR) (Xie *et al.*, 1995) but this motif could not be identified in any of the other Australian monocot-infecting mastreviruses. On the C-terminal portion of the *rep* the putative dNTP-binding domain was identified.

Virion-strand origin of replication

```

SSMV-1 TAATATTACCCGCCCC-----CACTCGCGAGGGCGACGTGTGCGCTCTACGTCGCCCCGAGC-----GGGCCGGCTTTTGT-----ACTG [120]
SSMV-2 TAATATTACCCTTGCCCCCGCGGGCCGTCCCAGGGCGACGTGTGCGCTCTACGTCGCCCCGAGGGAAAGTGATAGAGAAAGAAA---GGGAATCCTTTAAT-----TAAAGA [120]
DCSMV TAATATTACCGCCCATC-----TTTACCCGAGGGGCTTAC---GCTCAGCAGGCCCCGAGG-----CTTTGGCGCC-----CACTA [120]
PDSMV TAATATTACCGTCGCCCT-----GCTTTGCGAGGG-CCCAT---GCTCAACGGG-CCCGAGC-----GGTCCCGGCCCAT-----CCACTA [120]
BCSMV TAATATTACCGCCCATC-----TTTAGCCGAGGG-ACCAT---GCCCAACGGGTCCCGAGG-----CGCCCCGACCCCCCAAAC-----ACTA [120]
CSMV TAATATTACCGCCCCCCGG---CCCATGCGAGGG-CCCAT---GCTCAACGGGTCCCGAGC-----GGCTTTGGCTTT-----C-ACAT [120]
DDSMV TAATATTACCGCCCCCGG---CCATGCGAGGGCCCTAC---GCTCAGCAGGGCCCCGAGC-----GGCTTTACCTGCTTTGGTGTTCTTTAACTT [120]
PSMV TAATATTACCGCTGCCCTC---TTTACCCGAGGGCCCCAT---GCTCAACGGGGACCGAGG-----TGCTTTGCCCGGGCCTTATCAGC-----CCACTA [120]

                                     TATA box      Movement protein start codon

SSMV-1 GGCCGCATCTTTTCCCGGC---CTGAGTGCGCAACACGGGTGTTT-----GATCGGGACACCGCGTCTTTAAAGTCA-----ACATCTTTAGGTCCTTCGACGATGGAGGCGGG [240]
SSMV-2 AG---ACTTGGATAGCAC-----GTCTCTTTAAAGGGGGTGC-----GCTCTGAGT---CGGTTATGCTCTAATCC-GGCAACCCCTTTGGGTG--ATTAC----- [240]
DCSMV GG---GTTTGGCGCCCGC-----TGACGTATTGAGGTTGCCCCCG-CTCTAATTTGAGTGCCGTCTTTATAGACGG-GACCTCGCCGCCCTTCT--ATGGCGG-----AG- [240]
PDSMV AA---GATTGCTCCCGC-----ATTACGATAAATGGATTGACTGCCG-CTCTACTTTGAACGATGTCTATAAAAGACTG-CATCGCACTACGACGCC--ATGGCGG-----AG- [240]
BCSMV GG---GGTTGCACCCCGC-----AATATGATTTTCAATCTGACTGCCA-TCTTCCTTTGTTT--TGTCTTTATAGACTA-GATCGCACTAGTACGCT--ATGGCGG-----AG- [240]
CSMV GG---GCTTGTCCCGC-----GATGCGATCT---GCTCTGCCA-----TGCTTTGG---CGGCTTTATAGCCGTTCTCAGACCTTTGTTT--CCAATGC-----AG- [240]
DDSMV GG---ATTTTCGCAGTGC-----GATCTGCCATCTG-ACCGGTTTTGTCT-----TTTAAAGACCG-CTCCCCCTTTGTTTCGAATGCGA-----CTC [240]
PSMV AG---GGCTGTTTTCCGCAATCTGATTTTCGATTTTGTCTTTGACTGACATTTCTGCTTTGGATG-CGGCTTTAAATAGCCG-CATCGCACTACTACGCC--ATGGCGG-----AG- [240]

SSMV-1 CCATCTTCCATCGCAGCAGGGATTCCCATCGCCTTTGGCTTATTCCCA--GCCGAGCCCCAGCGGAGTCGGGAAC--GACTCCGCGTGGAGGACGCTCGTCCTGGTAT---TCACCATC [360]
SSMV-2 --ACCTCAAGT---CAG-----CATCAGG---GTGGTACGCA--GTCCGGATCCGTCCGAATCGGTAAC--GATTCCGCGTGGAGGGCGCTAGCTCTTGCTT---TCACCGTA [360]
DCSMV -TACCCTCAGT---CAG-----CCTTTGTTGGAGGTAGTGCGAT--TCCGCGTCAAGGCCCGGTGATAGCTTTGCTTCGACGTTGAAGGTAAGTCTTTGGGGCTCTTTGCTACG [360]
PDSMV -TACCCTCAGT---CTG-----CTTTAGTGGTAAGTGGTGCGAT--TCCACGGCGAAGCCAGGACGAGAGCTTTGCCTCGTCTTTGAAGTTCACTGCTTTATCTTTATTTGCAGCA [360]
BCSMV -TACCCTCAGT---CTG-----CTTTGCTTTTATCTGGTGCGAT--TCCACATCAAAGCAAGGACGAAGGCTTTGCCTCGTCTTTGAAGGTTACTGCTCTATCTTTGTTTCGACGCA [360]
CSMV -TACCAGGGGTACGAGCA-----GCTCAGT-AGATCTGGATCCGT--GGAGCAACCCAGCCCCGTGCTAGCTTTGCTTTCCCGGTGAAGGTGACAGCCCTCGTCTGTTTCGACGCG [360]
DDSMV CGACATTTGAT---AGTG-----CATATTTAGATCTGGATCTGTGGTTCAAACCAAGACGGGACA--AGCTTTGCTTTCCCGGTGAAGGTGACTGCCCTTGTCTGCATTTCTTCA [360]
PSMV -TACCCTCAGT---CTG-----CTTTGATTGTTTCTGGTGCGAT--TCCACGGCGAAGCCAGGACGAGAGCTTTGCCACGTCTTTGAAGGTCACTGCTTTATCTTTATTTGCAGCA [360]

SSMV-1 ACCGCAGTTGGTCTGGCGTGTTCAATTTGCGCTTTACCGTCTGTGTGTGAAGGACCTTGTTCTGTTGCTGAGGGCGAAGCGCTCAAGGACGGTGACGGAACGAGGTTTCGGCGGCACCCCG [480]
SSMV-2 ACTACGGTAACCTCTTGTGTTGCTGTTTCGGAGCTTGGAGGGCTTTGTCTGAAAGACTGCCTTCTGACTCTGCGGGCTAAACGCGAGCAAGACTACGACCGAAGTAGGATTCGGTCAGA----- [480]
DCSMV TTCATTGGGGCCGACAGTCTGTCAATTT--TTGTACAGGACGTGTCTATCAGACTGCATTACGCAGTACCGGCTTGGTTCTCTGGAGCAGTACCGTCAATTCGGGGCTTTGGGGGTAACCGG [480]
PDSMV TTTATTGCTGCTTGGCTCTTGTCAATC---ATTTACAAGACGTGCATTGCGGAATTCATTACGCAATACCGTTTATCCGGTTTAAGCAGTGTAATTCATCTGGCTTTGGGCGGACCGTT [480]
BCSMV TTCATTGCTGCTGCCATCTTGTGTTTC---CTATACAAGACTTGTCTTGCAAGATTGCTATACGCAATACCGGACTACCGGCTGAGCAGTACATCTTCTGCTGCTTTGGTTCGAACTCT [480]
CSMV ATTGTTGGAGCCTGTATCTTGTATTTC---TTGTACAAGACGTGCATTGCGGACTGCATAACGCAGTACCGGCTTACGGACTACGGCTGTACACTTCGGCTGGGTTCCGAGGTTGCGTTA [480]
DDSMV ATTTAGCGGCGCAATACTTGTATTTC---TTATACAAGACCTGTTTACAAGACTGCATAACGCAGTGGAGGCTTACGACATACGGCAGTCACTTTTCTGCTGTTTGGAGGTACTCAT [480]
PSMV TTTATTGCTGCTGCAGTCTTGACATTC---ATTTACAAGACTTGCATTGCGGATTCATTACGCAATTCGGTTTATCCGGTTTAAGCAGTGTAATTCATCTGGCTTTGGGCGGACCGCT [480]

```

Figure 2.1: continued on the following page.

Movement protein stop codon

SSMV-1 GCG---CGCCAA-----GACGGCGTTTCG-----TACCGGGAGTGGGGTTCCCGGGCTTGGA**TAG**TTTCCAC-----CAGGACCGGTTGCTTACATACGGGACTTC [600]
SSMV-2 -----CGCCGA-----GAAGGGATCCG-----GTAGGCGGAGGAG---TCACCCAGTACCCCC-----CCGGAGTCCCTGGG**TAA**CCAGGATCACCGC [600]
DCSMV GCGTTACCCAGGTCGTACGTGGAGACCTAGAAAGAC-AAGTATCTGTCCCC--GTAGGG-----ACTCTGGTGTTT-----TGCCTGTGGTTTCAGGAGCT--GC [600]
PDSMV GAGGATCCGCCAGCTGTCCCC--CGGACAGCGTCTG-AGGTTTCAATTCCA--GTTGCTG-----TTAGAAGTGGTATTCT-----CTCCTGTGATTTACAGGTGCT--TC [600]
BCSMV GAGGCTTCGACAGCTGTCCCG--CGGACAGCGTCTG-AGGTTTCTATTCCC--TTAGGG-----ATAGATCTGCTACTC-----CA--TCTTCTGTTTACCTACA-----TC [600]
CSMV CCCGTGACCTCTGC-----GCAAGCTAGTGCTGGTACCAGCACCCCTGTGTGTG-----TTCCCTGTGCTCCTCAGGTACAGGCGTCCGTGGATCTACCC-----TC [600]
DDSMV TTGGTGACCTCAG-----GCCAGCCTGAG--AGTAACCATACTCCCATTTGTGAGTTTACCTGATAGATCTGCAGTTT-----CAGTGCCCCCTGTTTCGTTACCTGTGGTT [600]
PSMV GAGGTTCCGCCAGCTGTCCCTACAAGGACAGCGTCTG-AGGTTTCAATTCCC--GTAAGGGAAG-----CTGTTTCTGCTCCTC-----CTTTGGTTTCTTCCG----- [600]

Coat protein start codon

SSMV-1 CGGCGAGCGCTCGAGCTGAGCCGA-----GCC**ATG**CCGTCGTCCACGCGTCTGCTGTGAGGCCTAAGCGTAAGAGGGTTGGGAAACAACCGTGGC-----CCAGAGAACTA [720]
SSMV-2 CGTAAAGACGCGCCGTGAAGCGGTCCAGAGAGGACGCT**ATG**CCTCCTCC-----CGTAAGAAGAAGAAGG--GTTCCGACTCTGGAGC-----CG-GAG----G [720]
DCSMV TGCAGTTGGCGGAGGCTGAGTC-----GTC**ATG**CCTGCTTC-----GTCGAAGAGGAAGCGTG--GGAGCAGTTCTGCTGG-----GAAGCG----G [720]
PDSMV CGGAATTCGCGAAATTGACGTTG-----GAA**ATG**CCGTCCTTC-----CTCAAAGAGGAAGAGGG--GGAGTACCTCGGGTAC-----TAAGCG----G [720]
BCSMV TGGTATTTCTATCTATT**TGA**TTT-----GTC**ATG**ACGTCCTTC-----CTCGAAGAGGAAGAGTG-----GATCTGGTAAGAC-----GAAGAC----G [720]
CSMV CGTAAGTAGGGTGTCT**ATGA**-----GTCCTGCCAGCTC-----ATGGAAGAGGAAGAGGC--CCTCTTCTTCTCCTCCGC-----TCAGGC----G [720]
DDSMV TGCAAACATCT**ATGA**-----GCCCTGCAGGCTC-----ATATAAGAGGAAGCG-----CTCCGCCGCTTCGTCTTTCAGCG----A [720]
PSMV -----TGGTTGGTGTT**TAG**TT-----TCC**ATG**GCAGCTTC-----GTCGAAGAGGAAGCGCG--GGAGTACTTCCGCAAC-----CAAGCG----G [720]

SSMV-1 TGGCAGAGGGGGATTACCC-----CACGAAACAGCAGGATCCGGTCGACGGGTCTCATCGATGAGACCCGTTTCGCCCTTTTCGTTGCAGTAATGAAGT--ATACATGGACACCCAAC [840]
SSMV-2 CTCCAGAAGCGTATACCAGCGAAGGTATTACGACCGAGGAGCTCAAGCGGACGAG-----CAGTCAGACGTCTCCTCCGCTTCAGTTCATCCAGTACAC [840]
DCSMV CGTAAGAAGC-----CGCGGTACACGA-----AGTGGACTGGT-----TCCC GTTCCAGCGCTAGCCAGGATGCACTGC--AGGTGCAGACCTTCCAA [840]
PDSMV CGTAAGAAGC-----CGCAACCACGA-----AGTGGACTGGAT-----CCCGCAGCGCATCTAGAGAGGCGCTGC--AAGTACAGACCTTCACC [840]
BCSMV CGCAAGAAGG-----CGCGTTACACGA-----AGTGGACA-----TCATCGAGAACAACGTGAGTGCAGACTCTCTGC--AGGTACAGACGTTCTG [840]
CSMV TCTAAGAAGCGC-----CGCG-----TGACAGGCTGCTGTTTACAGTTCTCTCGCTCGG-----GAGAACCTCTGC--AGGTGCAAGACTTTGTC [840]
DDSMV CGCCAAAACGCCGT-----CGCG-TCTATAAACAAGCAGTGAGTC-----GTCCTCTCTCAAGGAGAGAACCAGTGC--AGGTGCAAGATTTTACC [840]
PSMV CGCAAGAAGC-----CGCGGTACACGA-----AGTGGACC-----TCTGCCCGCAGCTCTAACAGAGATGCGCTGC--AGGTACAGACCTTCACC [840]

SSMV-1 GGGGCAGGAGTTTCAGG---TTGCTG-----CACCGGGTGCTGTCTATCTCATGACGAACTTGCCCCGGGGGAGCAGCGAAGACCAGCGACACACGGGGGAGACCTTGGCTTACAA [960]
SSMV-2 CTGGACGAGCAACGGTTCTCCGATAACCGTTGGTCCGAATGGGTACGTAGCTCTCCTGACTAGCTTCCCAAGGGGAAGCGACGAAGATAAGCGTATACAGGAGAGACTGTGACGTACAA [960]
DCSMV TATGCTGAGGATCAGGCATTTAATG-----CAGGAGGACGTGCGCTGTTGCTCAGCGCATTATACCCGCGGTTCTGCGAGAGAACCAGCGGAAGTCCAGGAGACCATTACGTACAA [960]
PDSMV TGGAGTGAAGACCAGGCGTTCAACG-----CCGGTGGCCGAGCGATCCTTCTCTCGGCCATACGCGCGGTTCTGGCGAGAACCAGCGCAAAATCCAGGAAACCATACGTACAA [960]
BCSMV TGGGCTGAGGATCAGTCCTTCAATA-----CTGGTGGAGGTTGACAGCTGCTAACCTCCTTACGCGTGGTTTCAGGAGAGAACCAACGCAAAATCCAGGAGACCATTACGTACAA [960]
CSMV TGGGATACAGATGTGGCTTTCAATA-----GGGAGGAGGATGCTACCTCCTCACTAGCTATGCTCGAGGCTCTGCCGAGAATCAGCGGAAGACCAGTGCAGACCATCACGTACAA [960]
DDSMV TGGGAGCAAGATTTCGCGTTCAATG-----CAGGTGGCTCCGCCCTACCTGCTTACAAGTTATGCCGCGGTTTCAGCTGAGAATCAGCGCAAGACCAGCGGAGACGATCACGTACAA [960]
PSMV TGGGGTGAAGATCAGGCGTTCAACG-----CCGAGGACGTGCGATACTTCTACGGCCTTACGCGTGGTTCTGCCGAGAACCAACGCAAAATCCAGGAGACCATTACATACAA [960]

Figure 2.1: continued on the following page.

SSMV-1 GCTGGGAATCGACCTAGAGGTACAGGTTGTGTGCTCAGCTCAGTTTGCCTATGCCAA-----CAAGAGTAC-----CCATGTCATGTGGCTGGTCTATGACGCACAGCCG [1080]
SSMV-2 GGTAGGCTTGGATCT-----G-----TTCTGCGTCAGAGACACAACAGATACAAGGAGATAGGATCTGCTATCCACTGCTGCTGGTTGGTCTACGACGCTCAACCA [1080]
DCSMV GGTTAGTGTTAGCCTTGGCGT-----TTCCGCATCTTCTAC-TGTACAGAAGTACTGCGTGAAGA--GCCAGCCG-ATATGCTGGCTCGGTACGATAAGACGCCT [1080]
PDSMV GGTTGCCCTCAACCTTGGTGT-----TTCCGCATCCGCTAC-GGTACTCAAGTACTGCTGCAGCA--GCCAGCCG-ATATGCTGGCTGGTGTATGACAAGACGCCT [1080]
BCSMV GGTTGCCCTCAACCTTGGCAT-----CTCCGCCTCCACTAC-CGTTCAAGAATATTGCCTGACCA--GCCATCCT-ATATGTTGGCTGGTCTACGATAAGACGCCT- [1080]
CSMV GGTGGCAGTTAACCTGGGGTG-----TGCTATCTCCGGGAC-GATGCAGCAATATTGCATCAGCT---CCCGACCG-GTCTGCTGGATTGTATACGACGCGGCCCC [1080]
DDSMV GGTGGCTATCAACCTTGGTGT-----TGCTATTAGTGGCAC-TATGCAACAATACTGTGTGTCGA--GCCGGCCA-GTGTGCTGGCTCGTTTACGATGCGGCACCG [1080]
PSMV GGTTAGTATTAACCTTGGTGT-----GTCAGCATCTGCTAC-CGTCTCAAGTACTGCTGCAAGA--GCCAGCCA-TTATGCTGGCTCGTATACGATAAGTCCCCG [1080]

SSMV-1 AGCGGTCTGCTT---CCGGCGACCTCTGATATCTTCGACTACGTTGAGGGGTTCCAGTATATCCACACGTGTGGAAGGTGAGGCGAGATCTGTGCCACCGGTACATTGTCAAACGGAAG [1200]
SSMV-2 ACCGGTACGATT---CCGGGGCTGGGAACCTATCTTCGACCTCGTTGATAAATTCCAGGAGTACCCGACGACATGGAAGGTAACCGGGGATATGGGGCACCGGTTTGTGATTAAACGTCGC [1200]
DCSMV ACTGGGATTGCTGACCTTGTCCCTCGGACATTTTCGATGTGCCGAGTGGATTGACCAATTGGCCTTCTACCTGGAAGGTCAAGCGAGAAGTCTCGCACCGCTTTGTGGTGAAACGGCGC [1200]
PDSMV ACTGGGATTACTGACCTGACCCCGTCAGACATCTTCGATGTGCCCTCGGGGTTACAGAAGTGGCCTTCGACCTGGAAGGTCAAACGCGAAGCGTCTCACCGCTTCGTGGTGAAACGGCGT [1200]
BCSMV --TGGGATTGCTGACCTGACTCCAAGTACATCTTTGATGTCCGACTGGGTTGAACAAGTGGCCTTCAACCTGGAAGGTCAAGCGCGAAGCATCTCACCGCTTCGTGGTGAAACGGCGC [1200]
CSMV ACTGGCTCTGCTG---TTACCCCGAAGGACATCTTCGGGTACCCGGAAGGATTAGTTAACTGGCCTACTACTTGAAGGTGGCCAGAGCGGTGTCCCACCGCTTCATAGTGAAGCGCCGA [1200]
DDSMV TCCGGCACGGCTG---TGACGGCTCAAGAAATTTTCGGATTCCCTGATGGTTTGAAGAAGTGGCCAACAATTGGAAGTGGCCAGATCGGTGTCCCACCGATTTCATAGTGAAGCGTCGG [1200]
PSMV ACTGGGATTACGGACCTGACCCCATCAGACATCTTCGATGTGCCCTCGGGGTTACAGAAGTGGCCTTCCACCTGGAAGGTCAAACGAGAAGCGTCTCATCGCTTCGTGCTCAAACGGCGT [1200]

SSMV-1 TGGATGATAAACCTCGAGACCAATGGAGCGTCCTTCGGGGTGGACTTCAGTAGCAGACCGGT-CACTGCTCAA-----AGTACCGGGCTAGCTTCCACAAGTTCGTTAAGCGGTTAGG [1320]
SSMV-2 TGGACCTTCACGCTCCAGTCGGATGGTCACCTGGGATCGAACGACTACA-----GCCGGGCTCCTGCAGCGCCCTGCAAGTACATGATTGCGTTCAACAAGTTCGTGAAGCGTCTAGG [1320]
DCSMV TGGCCGTTTACGTTGAGTTGCAACGGGAGTACCTTCACGGCGGATTACACGAAGCTGCCGTTGCCAATA-CAG-----ACAACCTGGTGTCCGTGAACCGGTTTCGCAAGGGATTAGG [1320]
PDSMV TGGCCGTTTAAAGTTGGAATGTAACGGCAGTACGTTCCAGAAGGACTACACGAACCTGCCTGTGTGCAATA-CGC-----AGAACCTGGTGTCCGTAAACGAGGTTTCGCAAGGGACTTGG [1320]
BCSMV TGGCCGTTTACGATGTCCGTCAATGGCTCTACGTTCTCTGTCAGATTATACGAAGCTGCCGTTGCCAATA-CAG-----ACAACCTGTGTACGATCAACAGGTTTCGCAAGGGACTTGG [1320]
CSMV TGGGTCTTCACCATGGAGTCCAACGGCTCGCGCTTCGACCGTACTACACCAACCTCCCGGC-TGCTATACCGC-----AGTCCCTTCCCGTTCTGAACAAGTTCGCGAAGCAGTTGGG [1320]
DDSMV TGGGTGTTTACACTGGAGTCCAACGGCTCCAATTTGCTACGGGGTACTCTAGTA-ACCCGTGTGCCATACCGC-----AATCCCTGCCGTTCTGAACAAGTTCGCAAGCAACTTGG [1320]
PSMV TGGCCGTTTCAGTTATCATGCAACGGGAGTACGTTCCAGAAGGACTACACAAACCTGCCTGTCTGTAAACA-CAG-----ACAACCTGGTGTCTGTAAACGAGGTTTCGCAAGGGACTTGG [1320]

SSMV-1 CGTACGAACTGAGTGGAAGAACTCCGACACGGGTGAGA--TCGGAGACATCCAGAGGGGAGCGTTGTACTTGGTGGTTGCTCCAGGCAACAACGTTCCGATAAACATTAGGGGGTACTTC [1440]
SSMV-2 TGTTCGGACGGAATGGAGGAACACAACGACGGGTGACA--TCGGAGACGTGTCTAGGGGTGCGTTGTACATCGTAATGGCTAGAGGCAATGCCTGGAGTTACGAAGTTAGGGGGCGTATA [1440]
DCSMV AGTGCGGACCGAGTGGAAGGATACGGTGTCTGCGGAGGCCTC--CGACATTAAGGGTGGAGCCCTTTACATAGTTCTTGCCCGGCTAACGGGGTTGTTTCCACCGCCCGCGGGTAAAT [1440]
PDSMV AGTGCGAACCGAGTGGAAGGACACAACGACGGCGGAGTCGTC--GGACATTAAGGGCGGAGCCCTGTACCTTGTGATAGCCCCTGCTAACGGGCTTGTGTTTACAGCCCGTGGTGTAAATC [1440]
BCSMV AGTGCGAACC GAATGGAAGGACACGGTTTCTGCTGACGCCTC--CGACATCAAGGGCGGAGCCCTGTACATAGTATTAGCCCCGCTAATGGGCTTGTATTACAGCTAGAGGTGTCATT [1440]
CSMV CGTGCGGACCGAGTGGAAGAAGC--TGAAGCGGAGACTTCGGCGACATAAAGAGCGGAGCTCTTTACCTAGTCATGGCTCCGGCTAACGGAGCTGTCTTTGTAGCCCGCGGCAATGTC [1440]
DDSMV CGTGCGGACCGAGTGGAAGAAGC--CGAGGGGGAGACTTCGGCGACATTAAGAGCGGCGCTCTTTACTTAGTGTGGCTCCGGCTAACGGATTAACTTTTGTAGCACGCGGAAATATC [1440]
PSMV AGTGCGAACCGAGTGGAAGGACACAACGACAGCGGACGCGTC--GGACATCAAGGGTGGAGCCCTCTACCTTGTGGTAGCCCCGCTAACGGGCTTGTGTTTACAGCCCGTGGTGTAAATC [1440]

Figure 2.1: continued on the following page.

Coat protein stop codon

SSMV-1 CGTTTGTATTTCAAGAGTGTGCGTAATCAGTAGGATTACCGTCTTTGTAATCGAAGATTAATAAG-----AAG-----CGAAGC-----TTTTATTTCAT [1560]
SSMV-2 AGAGTGTATTTCAAATCAGTCGGGAATCAGTGAATGAATAAG-----ACGTGTTTAA-----TAAAATGCG-----TCA- [1560]
DCSMV AAGGTGTACTTTAAGTCTGTGGGCAACCAGT-----AGCCAGTGT-AATGAGCCCTTGGGCGA-----GTATTAA-----TAAAAC--TCCAGTTTTATTAT [1560]
PDSMV AAGTGCTACTTTAAGTCTGTGGGAAATCAGT-----AGCCACTGTAATTGAGCCATTGGGCGATCTATGATCTGAAA-----TAAAATGGCACATTTTATTAT [1560]
BCSMV AAAGTGTACTTCAAGAGTGTGGGCAATCAGT-----AGCCCAATGT-AATGAGCCCATAGGGCGA-----TGAA-----TAAAATGGCACATTTTATTAT [1560]
CSMV CGCGTGTATTTAAGTCTGTTGGGAATCAGTGAATCCTCCAG-----ACTTCATTTTCAATAAAC---TGTGAGAGTTTGCTTGCCAAA-----CAACATA---ATTTCAATTCAT [1560]
DDSMV CGCATGTATTTAAATCCGTCGGTAATCAGTGAATCCTCCAG-----AGTTACCGTTCAT--GGTTTATGAATAATAAACACAGTTTGATGACAAGCTTGGCAAATTATTATTAAAATGAGGCA----- [1560]
PSMV AAAGTCTACTTTAAGTCAAGTGGGCAACCAGT-----AGCCCATTTGTAATGAGCCCATGGGCGA---ACTTATGAAA-----TAAAATCTTGCAATTTTATTTCAT [1560]

Replication associated protein stop codon

SSMV-1 ATAGCGTGCAGTGCACA-----GAGAAATTACAACACACA-CAATCGCAGCCG--AGGCTTAAGACGAGTCTCAGCGATTCCCTGCGATCATCTACATCAAAAAAACA [1680]
SSMV-2 -----GTATGACGCTTG-----ATTACATAACAGCACACTCGAACC GCCCGC--AGGCTTAAGACGAGTCTCGTGCAGTTCCCTGCGATACATGACATTAAAAAAGA-- [1680]
DCSMV TGCACTTGTGCAG-----TGGCGGTACGACACAGAAAATACAACATTAACTAAATTGGCGGCCAGATCGAAGGCGGCTAAGGGTTAGGACTCGAAA-----AAAAAACA [1680]
PDSMV CGCA-TTGCAGATGAACGCGTAGCGTTAC-----AAATTACAATACATA---GCGCAGCCTCGGGCTAAAGACCGAGTCTGA--AGCGACCTAGTACAATACCACCGGAAAAACA-- [1680]
BCSMV GTCA-TTATGACGAACG-----AGTACAACGAGAAAATTACATAATTGGTTTTGTGGGTGCGCAGGGGGAACCCG---GAGCACGCACCCAAAA-----ACACT--AAACACACA [1680]
CSMV AACGATGGCGCAGTATGCGC-----AATACATTTAAAGAAGG-----GCGGACAGGACAAAGCGGGCGGCTAAGGGAAGCCGCAAGGGGCAACACC-----ACA [1680]
DDSMV -----TTATGCCG-----ATTACAACATGG--TTTGTGTGTCGGGAGGGGGAACCCG---GACCACGCACCCAAA-----ACACTTAAATAAAAACA [1680]
PSMV CGCA-TTGAATGATGCGTAGCAGCAA---AGCAAATTACATAACTGG-TTTGTGGGTGCGCAGGGGGAACCCG---GAGCACGCACCCAAAA-----ACACT-AAATGAAACA [1680]

SSMV-1 CAC-----TTTATAATA-ATTCAAGCTGAGT-----CGCTGGCGTA-----AAGCCGGGTC-----CGGTATGAAGCTCTCTTCAGATGTCATGTAGTGGATC [1800]
SSMV-2 -----TTACATTTATGAAA-ATT-----ACA-----AAGCTTCATT-----TCTAA-TGAAGCTTTCTCCTTCCGACATATAATATATC [1800]
DCSMV CACCCAATCATATTAATGAAA-ATGCGCGGCGCGTGCCGACGGAGGAGTCT-----ACGCTGTAGCTGCGCCAAGGCTTTTGAAGAAAGTCTCTCCTTGGTACATGTAATGTACC [1800]
PDSMV -----ATATCTATCTGATA-ATGCTCTGCGCGTAGGTGTGTAACACCT-----ACGCCGTCGT-----TTCGA-AGAAACGGAACCCCTTGGTTCATGTAGAAGACG [1800]
BCSMV AACCAGAAATATCTATAGGATACATGGCCGGCGCGTGCCGATA-----GTC-----ACGCTGTAGCTGGCCCTAAGCTTCGA-AGAAGCTCTCCCTGCATACATGTAATGTATG [1800]
CSMV TCCCAAGAATGAGTTTGGAAAT-ATGCAGC-----CGCTGGCTTC-----ACGCTAATTC-----GAAGGAATGAAGTTTCCCTGGGAGAAGGTGATACACC [1800]
DDSMV CACCCAACATTACATTATTAAT-ATGCCGCCCTAG-----CGGAGAAGCT-----ATGCTTCGACGCTCCGACGAAGGAGGAATGAAAGACTCTCCTGCTTGAAGATAGTAGACC [1800]
PSMV CACCCAATAATCTATCTGAAA-ATGCCGACGAGGACTCGCGAAGGAGTCCGAGATTAAAGCTTCGA-----AGAAGCTTTCACCCGAGTACATGTAATGTACG [1800]

SSMV-1 TGCACGTTTCGCTTCGAAGTACGACACCTGCCCGGGTGTATATCGGCTAGCCAGTCTCTCGTCTCATTACCAGTATGATAGTAGGGATACCTCCCTTTATCAGTTTCTTCTTACCATAT [1920]
SSMV-2 GACATATTCCTTCAAATACGAACGCTGGCCGTCGTTTATGCTCTCATCCAGTCTGCTCGTCTCGTTAACTAGGATGATACTTGGTATTCCTCCTTTGATGAGTTTTTCTTGGCCGTAC [1920]
DCSMV TCGCAGTTAGCGTTAAACAGTCAAGCCTGACAGGCTGATTTGTTGGAGCCAGTCTTCGCTCTCGTTGGCGAGGATTATGATGGGACTCCGTTGGAGAGCAGCCGTTTTTGGCCGTAC [1920]
PDSMV ACGCAGTTGTCTGAAACCATGCAAGCTGTGATGGCTGCATGTAATGGGGCCAGTCTTCATCCTCGTTACGAGGATTATGGAAGGTATTCGTTAGGAAGAAGGCGTTTCTTCCCGTAC [1920]
BCSMV ATACAGTTTGTCTCAAACGACCAACCTGACTAGGTTGCATTTGTTTGGAGCCAGTCTTCGCTCTCGTTACGAGGATTATGATGGAAGTCCATTCTTGAGCATACGTCGTTTACCATAT [1920]
CSMV ACGGCGTTGGCGTAAACAGTCAAGCTGTTGGGTGACATCGACTGCAGCCAGTCTTCATCCTCGTTTACGAGGATAATACATGGAATCCCGTTGGGGATTCTTTTCTTCTTCCCGTAC [1920]
DDSMV ACACAGTTGGCGTAAACAGTCAAGCTGTTGGGTGACATCGACTGCAGCCAGTCTTCATCCTCGTTTACGAGGATTATGGAAGGTATCCCATTCGGGATAGTCTTCTTTTACCCTAT [1920]
PSMV ACTGCGTTAGCATTAACAGTCAAGCTGACTAGGCTGCATTTGTTGAAGCCAGTCTTCATCCTCGTTTACGAGGATTATGATGGAATGCCATTTGACAGCAGCCGTTCTTGGCCATAC [1920]

Figure 2.1: continued on the following page.

Rep A stop codon

SSMV-1 TTAGGATTTACAGTAAAGTCCTTCTGGCAGCCGATCAGAGCCTTCCAGTGAGGACAGAACTTGAAGGGGATGTCGTCGATCACATTGTATTGGGCCTCCTGGTC---GTATGTGGCCCAA [2040]
SSMV-2 TTTGGGTTGACTGTATATTCATGCTGAGAGCCAACCAAGGCCTTCCAATAAGGACAGAACTTGAACGGAATATCATCGATTACATTACTGGGCTTCGTTATTATAGT---TGGCGAAG [2040]
DCSMV TTTGGATTGACGGTGATGTCCCTCTGTGCTCCGACCAACCCCTTCCAACAAGGACGAATTTGAAGGGGATGTCATCAATTACATTGTATTACAGCGTCGGGAATGATGT---TGAGGAAG [2040]
PDSMV TTTGGATTACGGTGATGTCCCTCTGTGCCCCAACCAATCCCTTCCAACAAGGAACGAACCTTGAACGGGATGTCATCTATGACGTTGTATCTGGCAGTGGGAATGACGT---GGAGGAAG [2040]
BCSMV TTTGGATTTACGGTAATGTCAAACCTGGCATCCAACAAGCCCTTCCAACAAGGAACGAACCTTGAATGGTATGTCATCTATGACATTGTAAGTTGCGTTAGCAACTAGT---TGAGGAAG [2040]
CSMV TTGGGGTTTACCGTCAGGTCATACTGGCTGCCGACGAGTCCCTTCCAACAAGGACGAACCTTGAACGGGATGTCATCAATGATGTTGAACCTGGGCCTGGCAGTCCCATTCCTCTAGGAAG [2040]
DDSMV TTTGGATTACGGTTATGTGCAACTGGCTACCAACCAACCCCTTCCAACAAGGAACGAACCTTGAACGGAATGTCATCTATGACATTATAGATGGCATTTTTATTCCACTCTGTGAGGAAA [2040]
PSMV TTCGGATTTACTGTAATGTCCTTCTGGCTTCCGACAAGCCCTTCCAACAAGGAACAAATTTAAACGGAATGTCATCTATGACATTATATCTTGCATTTTGAATTACGCT---GGAGGAAG [2040]

Rep intron

SSMV-1 TCCACCTGCATATTGTAGTAGTTATGACGTCCTAACACCTGGCCCAAGATTTCTTACCGGTACGAGTTGGTCCGAGATGTAGAGGGATTTCCGGCGCCCTGC-----TCCGTCCTAG [2160]
SSMV-2 TCCAGGTTTACCTGCCAGTAGTTATGTTTACCTAGCTCTGGCCCAAGATGTCTTACCGGTACGATTTGGTCCGAGATGTAGAGCCAGATTTCTAGTTCCTGGATGGTCTGCAGAC [2160]
DCSMV TCGACGCTGTGTTGCCAGTAGTTATGCTTACCCAGACTTCTGGCCCAAGATGTCTTCCGGTACGACTTGGTCCGAGATGTAGAGGGATTTCTTCGTTTCCAGAGGTTTCTGCCCTGC [2160]
PDSMV TCCACGGAATGCTGCCAGTAGTTGTGACATCCAAGACTTCTTGCCCAAGAAGTCTTCCAGTTCTTGTGGTCCGAGATGTAGAGGCTTCTTCGCCCTTCCAGAGGACTTCTGTCTCTGC [2160]
BCSMV TCGACGCTGTGTTGCCAGTAGTTGTGAGCTCCAGACTTCTAGCCCAAGAAGTCTTCCAGTTCTTGTGGTCCGCAATGTAGAGTGATCTTCTTCGTTTATCAGGACGCTGTCTGTC [2160]
CSMV TCACTGAGTGCTGCCAGTAGTGATGAGTTCCAAGACTTCTGGCCCAAGAAGTCTTCCAGTTCTTGTGGGCCACAGATGTACAGGCTTCGTCCTCCGACTCCAGGCTCCAGACCTGG [2160]
DDSMV TCTACAGAGTGTGCCAGTAGTGGTGAAGTTCCAAGACTTCTGGCCCAAGAAGTCTTCCAGTTCTTGTGGTCCACAGATATATAAGGATCGTCTTCGATGATGCGGACCTCGTTCCTGG [2160]
PSMV TCCACGGAATGTTGCCAGTAGTTATGACTTCCGAGACTTCTTGCCCAAGAAGTCTTCTCGTTCTTGTGGACACATATGTAGAGGGAACGCTTTCGTTTCACTAGGACTTCTGTCTGTC [2160]

SSMV-1 TAACGTCAGCCATCCACTTTAAATCAGATTTCGGCGTCTCTGGCCGGA-----TGTAACAAGAGTACGCGAAGGGACTTACGATGTAGCACTGCAGGTTCTCTGATACCAGTCCAAG [2280]
SSMV-2 TAGCATCCATCAACCAGTTTAGGTCCCCGAAGCGTCAGCCTGCGGA-----TGTAAGAAGAGTATGCAGCAGGAGATACCTGGTA-----GAGTGTATTATCGAGCCAGTCTGTG [2280]
DCSMV TGGTGTATGCATCCATTGGAGGTCAAGCTTGGCTTGTCTCCGAAAGACCGGAGTGAAT-GCTGTAAGCATAGGGACTTACAGTGTA-----CAGTTCTGCTCGGAGCCATGCTCCG [2280]
PDSMV TCAATTTAGATATCCATTGGAGGTCAAGCTTGGCTTGTCTCCGTAAGTCCGTTGTGAAT-GGAATATGCATAGGGACTTACAGAATA-----TAATTCCTCTCGGAGCCATGCTCCG [2280]
BCSMV TAATATCAGACATCCATTGGAGGTCAAGCTTGGCTTGTCTCCGTAAGTCCGTTGTGAAT-GGAATATGCATAGGGACTTACAGAAAA-----TAATTCCTCTCGGAGCCATGCTCCA [2280]
CSMV TTAGGTCAAGACATCCATTGGAGGTCAAGCTTGGCTTGTCTCCGTAAGTCCGTTGTGAAT-GCTGAGGCTTGGAGACTTACAGTATA-----CAGTCTGCTGGAGCCATTCTCTCT [2280]
DDSMV TTACATCAGACATCCATTGGAGGTCAAGCTTGGCTTGTCTCCGTAAGTCCGTTGTGAAT-GCTGAGGCTTGGAGACTTACAGTGAA-----GAACTCGGTATCGAGCCACTCTCCA [2280]
PSMV TTAGATCAGACATCCATTGGAGGTCAAGCTTGGCTTGTCTCCGTAAGTCCGTTGTGAAT-GGAATACGCATAAGGACTTACAGTGTA-----TAATTCATCTCGGAGCCATGCTCCG [2280]

SSMV-1 AGGTTCTCATGACAT-----GTGAGATCC--GTAGTCTGGTACTGGCTCTGGTACTGGGGAGTGATGTCAGGGAAGAGCTTGGATGCAAGATATTGAACTGGGCTAGTCTAGTAGCC [2400]
SSMV-2 ATTCTCTCATACAC-----ATGAGATTT--TCTGTAGGGAAGGGCTTGTGTATTCCGGAATCACGTACGGAAGAGCTTGAACGCGAATATTGAACTGCTGGAGTTTAGTAGCC [2400]
DCSMV AGCACTGGATGTTCTTCTTGGCTTGGCATGT-----TGAATGGGCTGACGTAGGGCGGAGGAGTAGAGGGGAACAGCCGTTTCGGCTGAGTATTCAAAGTTTGGCAACCTTGTGGCC [2400]
PDSMV AGCACCTGATGTTCTGCCTGTGAAGGCATCC-----CGAAGGGATCAACATACGGAGGAGGGAGGACAGGCGCTTGCCTGAGTATTCGAACTGTTGCAACCTTGTGGCC [2400]
BCSMV ATCACCAGGATGTTTATCTGTGCTGGCATCC-----CGAAGGGATCCATATATGGCGTAGGAACCGAAGGGAATAAGCTCTCTGCGGAGTATTGGAATTGTTGCAACCTCGTTGCC [2400]
CSMV ATGACCGGATGGTCG-----GACATGTCCCTTGACGCGTATGGCGGAGTAGGTTTGAAGGGGTCAGGGAACAGGCGATTGCACTGTATTGGAATTGCTGTAGGCGTACCGCC [2400]
DDSMV ATGACCGGATGGTCA-----GACATGTGCTTGGAGACGAAGTGTGGCTGGTATTGTATCGGAGCTTCCGGGAAGAGCGCTTGGCGCTGTATTCAAACCTGTTGGAGACGGATTGCC [2400]
PSMV ATAACAGGATGTGTATCTGCGATGGCATCC-----CGAAGGGATCCACATACGGTGGAGGAGTGAAGGGAACAGCGATTTCAGCTGAGAACTGAACTGCTGCAACCTTGTGGCC [2400]

Figure 2.1: continued on the following page.

SSMV-1 CACTCAAACGGGAAGGACTTTCGGACCATCCCGAGATAATCGTCCCGGGTGGTAGCGGTTTCGGAGGATGTACCGCATGCGCTCGTCCCTTGACGGCCGAGTC-GCTGCGCCTCCTATCTGT [2520]
SSMV-2 CAGTCAAATGGGAAGGCTTTTCTGACCATTTGAGAGATAACTGGCTCGGTCAAGTGTGATGATTGAGCGCATGACTGAGTCCTTTGCTTTGGACTC-AGTGTCTTTGGAGTCCT [2520]
DCSMV CAGTCGAATGGGAAGGCTTCTCTTACCATTGAGAGATACTCCTCCTTGCTCTTTGGAGCTTCTGATGATATCCGCCATTTTGGCGTCCCTTGTGTTGACGC-GTCGATCTT-----GC [2520]
PDSMV CAGTCGAATGGAAATCTTTTGGACCATCCCAAGATAGTCTCAGTGATGTTGATTGAGAAATAATAGACGCCATTTTGGCGTCCCTCGTGGAAGGAGC-GTCAGCCTTCTTCTTCTT [2520]
BCSMV CAATCGAATGGGAACGTATTCCTGACCATCGAGAGATAGTCTCCTTGCGAGGTACTGCTCTTGATTATCTCCGCCATTTTGGCGTCTTTAGTTGATGGAGC-ATCAACCTTCTTCT---- [2520]
CSMV CATTCGAAAGGGAACCTCTTTCGAACCATGCTCAGATATTCGTCCTTGAGCGTGGCGTTTGCCATGATTGTTTCATGGTCTTGTCCCTAGAGGCCGACTGGGTGCGACTCCTGTT---- [2520]
DDSMV CAGTCAAACGGAAATGCTTTCCGTATCATGGACAAATATTCGTCCTGCTGTGTGGCGTTAGCCATGATTGTTTCATTTTTTGTGACGGGTAAAGGAAGCTAGCTTCCTTTTCTCGAT [2520]
PSMV CAGTCGAATGGGAAGGATTTCCTTACCATCGAGAGATAGTCTCCTTGTTGTGGAACTTTAAATGATTTCGCCCATTTTGCATCCTTCGTAGAAGGTGC-GTCCGCTTCTTCT---- [2520]

SSMV-1 GTGTTTGGGAGGTCTTCCGCCAGGGGCGACAAACTTCCCCCTGGCTGACTGGCTAACAGGCTCTTTGAGGATATACTCTTTGACCTTAGTAGCGCTTCTAACTGTTTGGATATTTGGGTG [2640]
SSMV-2 GCCTTTGAGAGGAATG-----AAGGTTCTCTCTCGGCCCTTTGAGACTGGATTTTTGAGACAATAGTCTCGGACGCTTGTAGGAGATTTTGCAGTTTGGATGTTAGGATG [2640]
DCSMV GTTTGCGGACGTCTCTCTCCGTG-----TACGCCCCCTCTCGGCGTACGATACCGGAGATTTCTTGCAATGCAAGGACTTTGTGGGGCATCCTTGCCTTTGTACGTTGGGATG [2640]
PDSMV CTTCTTCGCTTCGTCTCTG-----AAAGCCCCGTGCTCAACGAAGCATACGGGGCCTTTCTTGATGTACGCAAGTACCTTGTGAGGTACCCGCGCTTCTGCACGTTAGGATG [2640]
BCSMV -TCTTCTGCTTGGTGCTTCTG-----GAAAACCCCCGTTTCTACGAAACATAAGGGGTTTTTTTTGATGTATGCTAACACCTTGTGTGGCATCTTGCATTCTGCACGTTAGGATG [2640]
CSMV GACCTTGGGTTTCAGG-----AATTTTCCAAATCCCAGCTACTCTCAGGATGTTTCATACAGTATTTTCAAGTACTGGCTGGCTGCCTTGCAGCTTGGATATTTGGATG [2640]
DDSMV GCCTTGGGCTTTACG-----AAGACCCCGTCTTCGTAGAAATCTGCCGATTTTTCTGACAGTATTCAGAGTTTTCTCTGGATTTCTTGGCGCTGAATGTTAGGATG [2640]
PSMV -TTCTCGGACGCTTGATTTCCTGA-----AATACCCCATATTCGCTTCGGATATTGGGGATTTCTTGCAATGATAATGCCCTTCTTGGGCATTCTTGGATTTTGAACATTTGGATG [2640]

SSMV-1 ATTTCCACCCAAATCAGCAAAAGAAGAGTCGCGAGACCGATATTGATCAGTAAGCTGCACAAGGCAATGAACATGGAAGCCGGAGTCTGATGCAGCTCACGAACAGAAAGAATATACAG [2760]
SSMV-2 ATAATCAAGGATATTAAGTACTGCTATCATTTGGTACTGAATGCCTTATCAAGCTGAAATAGACAGTGAAGATGGTAATCACCGTCGCTGTGATGCTCACGAGTGACGAGGCAGTACTT [2760]
DCSMV GTATTCTTCGAGGTGGAAGAAGCGTGACGCGTAGTGCGTACGTGTTTCTTGCAATGTCACGAAGCAATGCAGGTGGAAGGTGCGCTTGGTGTAGTTCTCTTGCCACGTATACATACGT [2760]
PDSMV AAATTCCTTCGACATCGAAGAATTTAGCACTCTTAGTGCGAATGTATTTCTGCACTGTACTAGACAGTGCAAAATGGTAAGAGCCATCCTTGTGCTCTTCCCTTGCACGTATGAATACGT [2760]
BCSMV AAATTCCTTCGACGTGGAAGAATTTAGCACTCTTGGTGCGAATATTTCTTGCAATGTACCAGACAATGCAAAATGGTAAGATCCATCTTGATGTTCTTCTCTTGCCACGTAGACATACGT [2760]
CSMV AAATTCATCCAGGTCAAAGTATTTTGGAGAGGTGGTCTGAAATTTGCTTCGAGTTGGACGAAGGCATGTAAGTGGGGCTCACCGTCAGCATGGAATTCCTGGAATGTAATATAAATT [2760]
DDSMV GAATTCAAAGAAATCAAAGTATTTGGCGGATGTGGTTCGGAAGAACTTTGTCAAATGAAGGAACGCATGTAATGCGGTTCCCATCTTGATGTAGCTCTTGAGCGATATACATATACTG [2760]
PSMV AAATTCCTTTGATGTGGAAGAATTTGCACTTGTAGTGCGCACGTATTTTGGAGCATTGTATTATACAATGCAGATGGTGAGATCCATCTTTATGTGCTCTTGTGCCACGTAGATATACGT [2760]

SSMV-1 AGGCTTATGAGACGAAAATTTGTCCCAGAGAGCATCTGTGATTAGGGCTGGATCGATCTCACACCGGGAGTAGGTGAGGAAGATGTTCTTCCCTCGGAAGTGAATCC----- [2880]
SSMV-2 CGGACCGTACTTGCGAATTTAGAATAGAGATGTTGACAACATCTCTAGGCTCCAGAGTGCACTTCGGATATGTAAGAAATGCGCTACGAGCTCTGAATCTGAAGTTT-GCTGGGGATG [2880]
DCSMV AGGCTCAAACCTTGCGAAGTTTGTGCGTGATGTACGCAACGCTTCCTGTGGATCAAGCAAGCACTTGCTGTATGTTAGGAAGATGTTCTTGGCCCTCACCTCGAAGCTT-GCTTCGGCAG [2880]
PDSMV CGGGTCCCACTTCTTAAGAAGTCTGGAGAGGCGTTCAAGCATGAACGACGGCTCCAGGTGGCACTTACTGTATGTGAGGAATACATTCCTGCTTCTTACCTCGAAGCAA-GCTTCGACCG [2880]
BCSMV GGGTTCGAACCTTCAAGAAGGCTTGAGAGGTGTTCTTGCAATGAACACGGATCAAGGTGGCACTTGCTGTACGTTAAGAAAATGTTCTTGATCTTACCTCGAAGCAT-GCTGCGACCG [2880]
CSMV ACATTCTTATTTTGTAGACGGTCAGCAATTTTCTGACAGCCTTTCGGGACTGATAGGACACCTGGGATAGGTGAGGAAGACATGTTAGTCTCTCAGGGAGAAG-----GCCTT [2880]
DDSMV TATATTATACCGCTTACAAAGTTTGTGAGAAACGAACAGCATCCTTTGGAGGGATTGGGCACCTTGGGTATGTGAGGAAGACATGTTGAGAGCGGACGTTGAAGCTTAGCGGCAGACT [2880]
PSMV GGGACAGTATTTCTCAATAAAGAAGAAATATATTCCAGGAGGAATACAGCCGTGAGGTTGCACCTTACTGTATGTGAGGAACAGGTTCTCGCTCTCACCTTGAAGCAT-GAAGAGGAGG [2880]

Figure 2.1: continued on the following page.

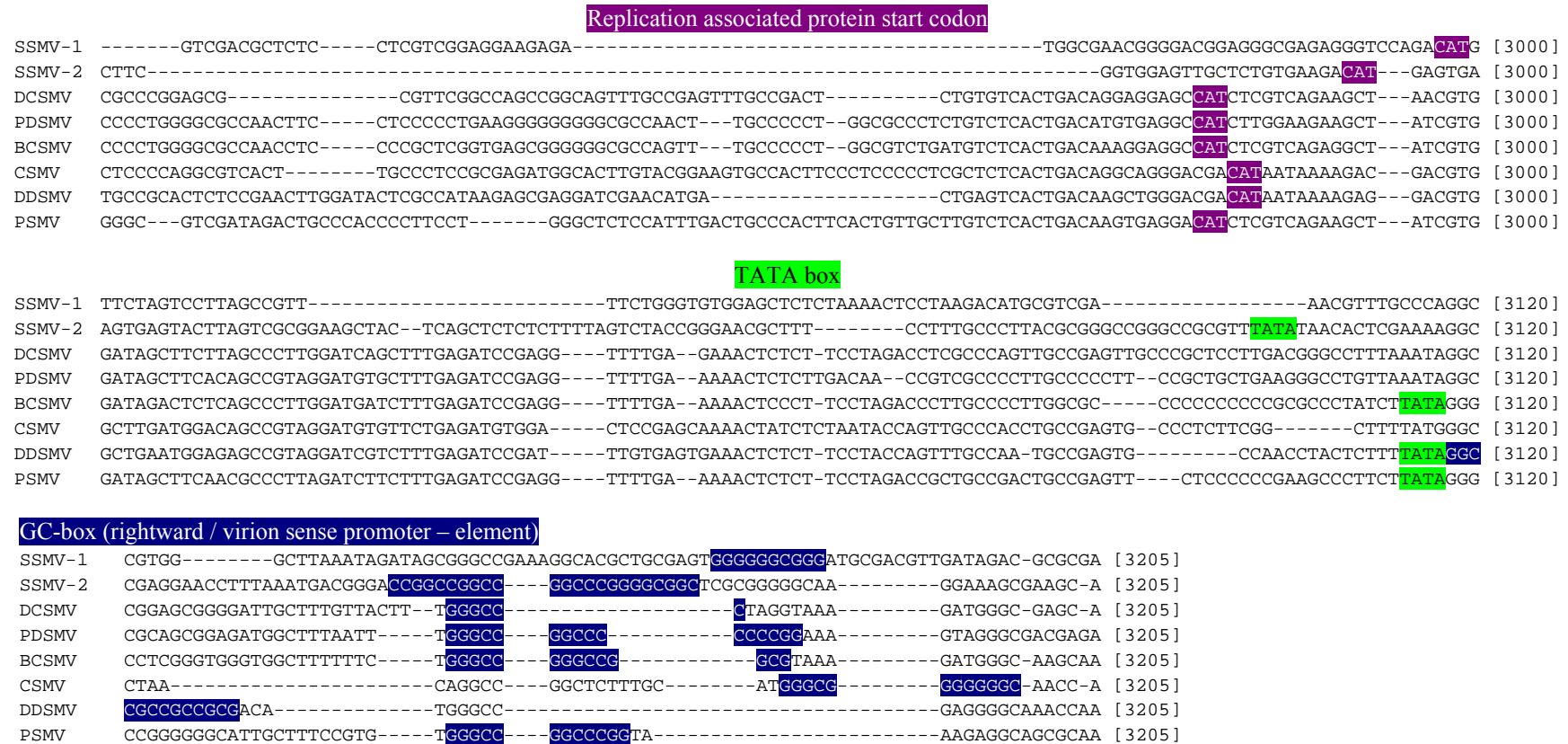


Figure 2.1: Nucleotide sequence annotation of a representative from each Australian monocot-infecting mastrevirus species aligned with MUSCLE.

Replication-associated protein

	Rolling circle replication motifs (Koonin & Ilyina, 1992)	DNA binding domain	Rep catalytic domain	GRS motif (Nash <i>et al.</i> , 2011)	
SSMV-1	M-SGSPRPSPFAISSSDEESVDG-----FHFGRKNI	FLTYSRCEIDPALITDALWKFSSHKPLYILSVRELHQDGS	FVHCLVQLTDQYRSRSDSSFADLGGNHPNIQTVRSATKVKE	YILKEPVQSARG	[140]
SSMV-2	M-SSQSNSTE-----ASPANFRFRARSA	FLTYPKCTLEPRDVVEHLYSKFRKYGPKYCLVIREHSDGDY	HLHCLFQLDKAFSTNDSSTFNILDYHPNIQTAKSPTNVR	FYCLKNPVSKAERG	[140]
DCSMV	M-APPVSDTESANSANCLRAER-----APGAEEASFVRAKNI	FLTYSKCLLLDPQEAALRDITHKLKRFPTYVYVARELHQDGT	HLHCFVQCKKHVRTTRARFFDLEEYHPNVQ	NARMPHKVLA	YCKKSPVSYAEEG [140]
PDMSV	M-ASHVSETE--GARGQVGAPPLQGEVVGAPGAVEACFEVRSRN	FLTYSKCHLEPSFMLERLSRLKKWDPTYSYVAREEHKDGSI	HLHCLVQCRKYIRTKSAKFFDVEEFHPNVQ	NARVPHKVL	YIKKGPVCFVEHG [140]
BCSMV	M-ASFVSETS--DARGQTGAPRSPSGEVGAPGAVAAACFEVRSRN	FLTYSKCHLDLPVFMQEHLSLLRRFEPTYVYVAREEHQDGS	HLHCLVQCKKYVRTKSAKFFDVEEFHPNVQ	NARMPHKVLA	YIKKNPLCFVETG [140]
CSMV	MSSLPVSESEGEESGTSVQVPSRGGQV--TPGE--KAFSLRTKHV	FLTYPRCPISPEEAGQKIADRLKNNKCNYYISREFHADGEP	HLHAFVQLEANFRTTSPKYFDLDEFHPNIQAARQPASTL	YCMKH	PESSEWFG [140]
DDSMV	MSSQLVSDSVMFPDRSYGEYPSSE-----SAASLPLSNVRSQH	FLTYPRCPIPPDKAGSFLKKLCRYNIQYMYIAQELHQDGE	HLHAFVQFDKVFRTTSAKYFDFFEHPNIQAARNPEKTL	YCKKN	PADFYEDG [140]
PSMV	M-SSLVSETSNSEVGSQMESPGRGGQSIDAPSS--SCFKVARN	FLTYSKCNLTAVFLLEYISLLKKYCYTYIYVAQEAHKDGS	HLHCIIQCSKYVRTTSAKFFDIKEFHPNVQ	NPRMPKKALS	YCKKSPISEAEYG [140]
		Oligomerisation domain (Horváth <i>et al.</i> , 1998)	LxCxE motif	dNTP-binding domain (Xie <i>et al.</i> , 1995)	
SSMV-1	KFVAPGGRPPKHTDRRRSDSAVKDERMRYILRTATTRDDYLGMRKSF	FEWATRLAQFEY	SASKLFPDITPQYQSQYQTTDLTCHENLLDWYQENLCYIVDGARRK	SLYICGPTRTGKKS	SWARVLGRHNYNMQVDW [280]
SSMV-2	TFI-----PLKGRTPKNTESKAKDSVMRSIINTSTDRASYLSMRKAF	FPFDWATRLQQFEY	SASKLFPDVIPEYTSFPFTEMLMCNERITDWLDNTLYQSADHPTRKS	GLYICGPNRTGKTS	WARSLGKHNYWQMNLD [280]
DCSMV	AYT-----ERDVRKRKIDASTTKDAKMAEIIRSSKSKEEYLSMRKTF	FPFDWATRLQNFY	SAERLFPSTPPPYVSPFNMPSQEHPVLGAWLRAELYTQGRNPAERRK	SLYICGPRSTGKTS	WARSLGKHNYWQHSVDF [280]
PDMSV	AFKD-----EAKKKKKKADAPSTKDAKMAIISQSTSRDYLGMVKKE	FPFDWATRLQQFEY	SAQALFPCLPPPYVDFGMPSPAETHQVLGAWLREELYSQDRSPAERRR	SLYICGPTRTGKTS	WARSLGCHNYWQHSVDF [280]
BCSMV	VFQA-----STQKQKKKVDAPSTKDAKMAEIKSSSTCKEDYLSMRN	TFPFDWATRLQQFQY	SAESLFPSPVTPYMDPFGMPAQDEHPVIGAWLQAELEFS--DRRDPERRR	SLYICGPTRTGKTS	WARSLGAHNYWQHSVDF [280]
CSMV	KFL-----KPKV-NRSPQTQSASRDKTMQIMANATSRDEYLSMRK	SFPFEWAVRLQQFQY	SANALFPDPPQYTSAPYASRDMSDHPVIGEWLQQELYT--VMSPGVRRR	SLYICGPTRTGKTS	WARSLGTHHYWQHSVDF [280]
DDSMV	VFV-----KPKASRKRKLASFTRDKMKQIMANATSRDEYLSMRK	AFPFDAIRLQQFEY	SAKALFPEAPIQYQPFVSNMDSMDHPVIGEWLDEFTT--ERGPVRRR	SLYICGPTRTGKTS	WARSLGTHHYWQHSVDF [280]
PSMV	VFQE-----IKRPRKKKADAPSTKDAKMAEIKSSSTCKEDYLSMR	KSFPFDWATRLQQFQY	SAESLFPSTPPPYVDFGMPSQDTHPVIGAWLDELTYT--DRSPTEERR	SLYICGPTRTGKTS	WARSLGSHNYWQHSVDF [280]
SSMV-1	AT-YDQEAQYNVIDDIPFK	FCPHWKALIGCQKDFTVNPKYGKKKLIKGGIPTIILVNEDEDWLADMT	PGQVSFYFANVQIHYMTSEESFIPDPALRQRLSLNYYKVCFFLM		[391]
SSMV-2	AN-YNNEAQYNVIDDIPFK	FCPYWKALVGSQHEYTVPNPKYGKKKLIKGGIPSIILVNEDEDDWMRAMNDG	QRSYFEGNMSIYYMSEGESFIRNEAL-----		[391]
DCSMV	LN-IIPDAEYNVIDDIPFK	FVPCWKGLVGAQRDITVNPYKGRKRLLSNGVPCIIILANEDEDWLQMQP	QGADWFNANCEVHYMYQGETFFKSLGAATA-----		[391]
PDMSV	LH-VIPTARYNVIDDIPFK	FVPCWKGI VGAQRDITVNPYKGRKRLLPNGIPSIILVNEDEDPQYMQPS	QAQAWFQDNCVVYMYNGGFRFFETTA-----		[391]
BCSMV	LN-LVANATYNVIDDIPFK	FVPCWKGLVGCQFDITVNPYKGRKRLKNGVPSIILVNEDEDLQMQPS	QVGFETNCIIHYMYAGESFFEA-----		[391]
CSMV	LEWNCQAQFNVIDDIPFK	FVPCWKGLVGSQYDLTVNPKYGKKKIPNGIPCIILVNEDEDLQMSMT	QQVDFHGNVAVVYHLLPGETFIPSE-----		[391]
DDSMV	LTEWKNKAIYNVIDDIPFK	FVPCWKGLVGSQFDITVNPYKGRKKTIPNGIPSIILANEDEDWLQMTSP	QADWFGNCCVVYVLQAGESFIPSSDVEA-----		[391]
PSMV	LH-VIQNARYNVIDDIPFK	FVPCWKGLVGSQKIDITVNPYKGRKRLLSNGIPCIILVNEDEDLQMQP	SQADWFNANAVVHYMYSGESFFFEAL-----		[391]

Figure 2.2: Amino acid annotation of a representative from each Australian monocot-infecting mastrevirus species aligned with MUSCLE. Within the N-terminal portion of the Rep of all the viral isolates we identified the four motifs that are conserved in all geminiviruses and which are usually also found in the replication initiator proteins of other ssDNA viruses. Motif I (amino acids FLTYx), a double stranded DNA-binding domain, binds iterated sequence elements (called iterons) in the LIR near the *v-ori*. Motif II (amino acids H[V/L]H[C/A][L/F]xQ) binds divalent ions (Mn^{2+} and Mg^{2+}) and is possibly involved in DNA cleavage (Argüello-Astorga & Ruiz-Medrano, 2001; Gutierrez, 1999; Orozco & Hanley-Bowdoin, 1998). Motif III (amino acids YxxKx) catalyses DNA cleavage at the *v-ori* of replication (Laufs *et al.*, 1995; Orozco & Hanley-Bowdoin, 1996). The fourth motif, named geminivirus Rep sequence (GRS; H[L/V]H[C/A]xxQ), is vital for the initiation of RCR (Nash *et al.*, 2011). We identified a conserved LxCxE motif that likely binds the host's Retinoblastoma related protein (pRBR; (Xie *et al.*, 1995) in the two highly divergent Australian monocot-infecting mastreviruses (SSMV-1 and SSMV-2) but not in any of the other Australian monocot-infecting mastreviruses. On the C-terminal portion of the Rep we identified a likely dNTP-binding domain.

2.4.2 Classification of novel Australian monocot-infecting mastreviruses

Forty-one full length mastrevirus genomes were isolated from 40 symptomatic grass samples from the *Poaceae* species; *Paspalum dilatatum*, *Digitaria ciliaris*, *Ehrharta erecta*, *Chloris gayana*, *Panicum* sp., *Sporobolus* sp., *Triticum aestivum* and *Eriochloa polystachya*. Our analyses revealed that of the 41 isolates, two isolates from a single *Sporobolus* plant are highly divergent and unlike previously described Australian monocot-infecting mastreviruses (Fig. 2.3A and B). Pairwise distance calculations (with pairwise deletion of gaps) revealed that these two unique mastreviruses share ~63.9% pairwise identity with one another and less than 62.5% identity with all other available mastrevirus sequences (Fig. 2.4). Based on these very low degrees of similarity, strong phylogenetic support for separation of these sequences from all the other known Australian mastreviruses (Fig. 2.3A) and the International Committee on the Taxonomy of Viruses (ICTV) sanctioned <75% mastrevirus species demarcation threshold (Brown *et al.*, 2011) we propose that these isolates represent new species which we have tentatively named *Sporobolus striate mosaic virus 1* (SSMV-1) and *Sporobolus striate mosaic virus 2* (SSMV-2).

The remaining 39 isolates cluster with previously described Australian monocot-infecting mastrevirus species (BCSMV, CSMV, DDSMV and PSMV; Fig. 2.3B). Pairwise distance comparisons indicated that twelve of these are CSMV isolates (>96.8% identity to the known CSMV) and seventeen are PSMV isolates (>92.3% identity to the known PSMV) (Fig. 2.5).

Six isolates, all from *Paspalum dilatatum*, share ~68% similarity with PSMV (calculated according to ICTV specifications but 78% identity when calculated from pairwise alignments with gaps excluded as a fifth nucleotide state), their nearest currently described relative, and therefore should probably be assigned to a novel species which we have tentatively named *Paspalum dilatatum striate mosaic virus* (PDSMV). The remaining four isolates, all from *Digitaria ciliaris*, share <73% pairwise identity with all other mastrevirus species available and likely also represent a new species which we have tentatively named *Digitaria ciliaris striate mosaic virus* (DCSMV; Fig. 2.3B and Fig. 2.5).

Our analysis of the proportion of pairwise identities between 527 monocot-infecting mastreviruses available in GenBank and the 41 from this study (161013 pairwise identities compared) reveals a logical demarcation pattern for strains (80–94%) and genotype / variant (95–100%; Fig. 2.6). Despite the larger number of sequences compared, our new pairwise distribution analysis is consistent with previous analysis carried out for assigning strains and genotype/variants to mastreviruses (Hadfield *et al.*, 2012; Martin *et al.*, 2001; Varsani *et al.*, 2009a). Based on these pairwise identities, the CSMV, DCSMV and PSMV isolates could be further split into strain and genotype/variant groupings (Fig. 2.6; Table 2.1). Amongst the DCSMV and PSMV isolates, we identified two strains (DCSMV-A and -B and PSMV-A and -B) and further classified the PSMV isolates into eight genotype/variant groupings (PSMV-A₁ to A₆ and PSMV-B₁ and -B₂). Similarly, we classified isolates within the single known CSMV and PDSMV strains into six (CSMV-A₁–A₆) and three (PDSMV-A₁–A₃) genotype/variant groupings, respectively.

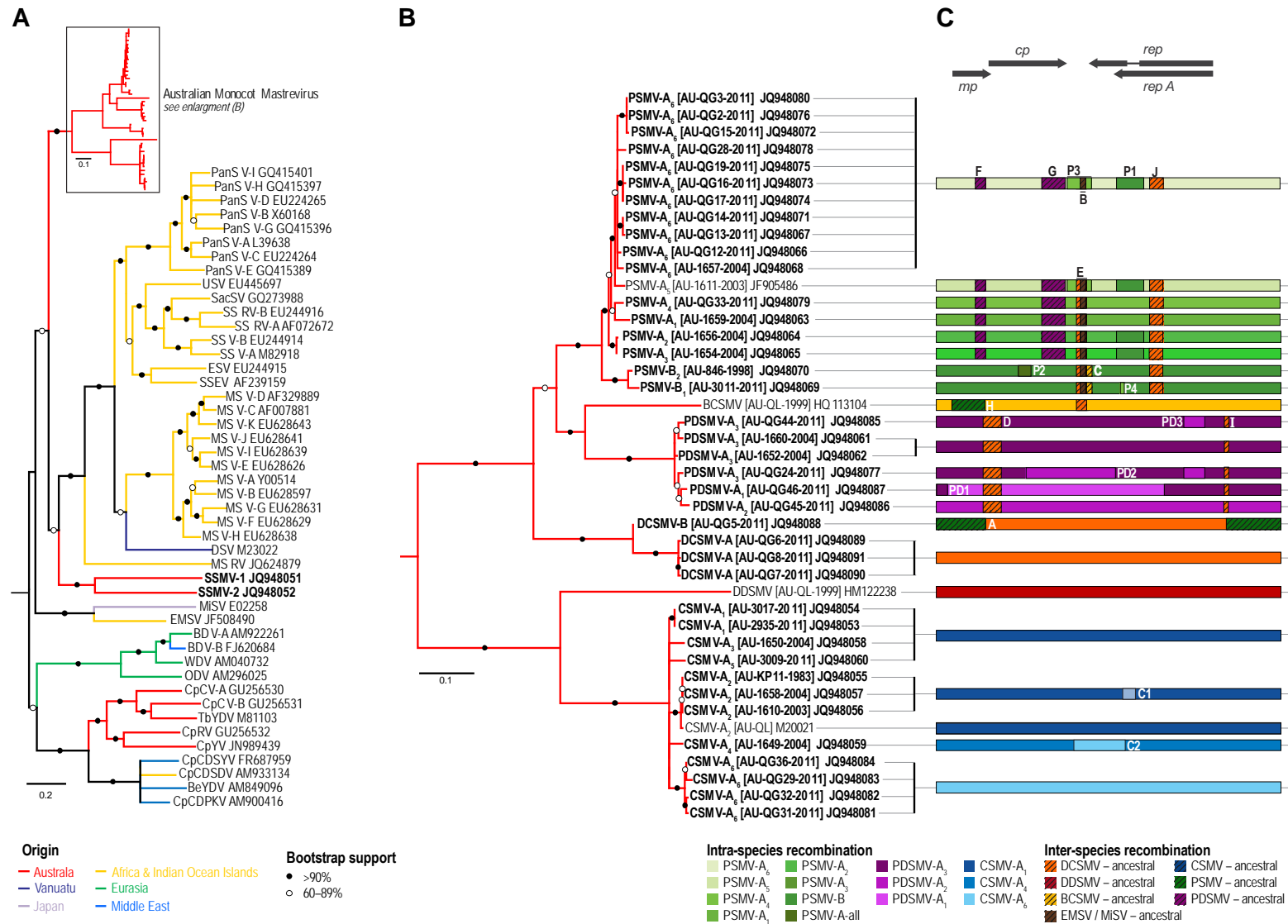


Figure 2.3: See following page for figure legend.

Figure 2.3: (A) Maximum likelihood phylogenetic tree of representative mastrevirus full genomes from each species and major strain grouping. Branches are coloured according to either known (terminal branches) or likely (internal branches) regions of origin. Bootstrap support of branches is indicated by open and closed circles, branches with less than 60% bootstrap support have been collapsed. Viral isolates found in this study are indicated in bold. (B) Enlargement of the Australian monocot-infecting mastrevirus branch of the maximum likelihood tree of viral isolates recovered in this study are indicated in bold. (C) Cartoon depicting recombination events amongst Australian monocot-infecting mastrevirus isolates. Inter-species recombination events are indicated by a number and intra-species recombination events are indicated by a letter followed by a number. The colouring within the cartoons corresponds to the likely origins of recombinationally derived genome fragments. Genome map shows position of the *mp* (movement protein), *cp* (coat protein), *repA* (replication associated A protein) and *rep* (replication-associated protein) genes.

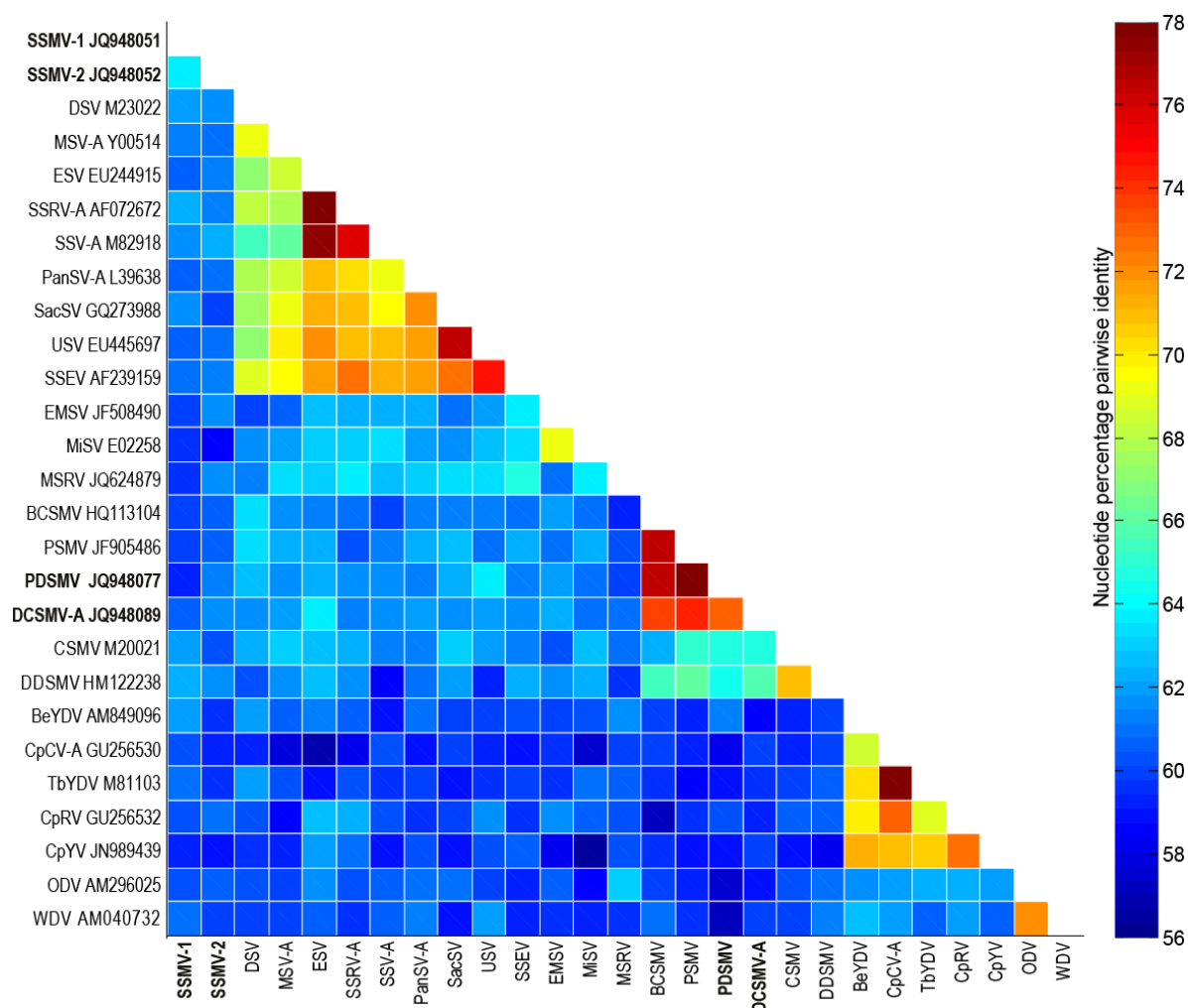


Figure 2.4: Two-dimensional percentage pairwise nucleotide identity plot (calculated with pairwise deletion of gaps) of full mastrevirus genomes, with a single representative from each species (tentative new species of Australian monocot-infecting mastrevirus are highlighted in bold).

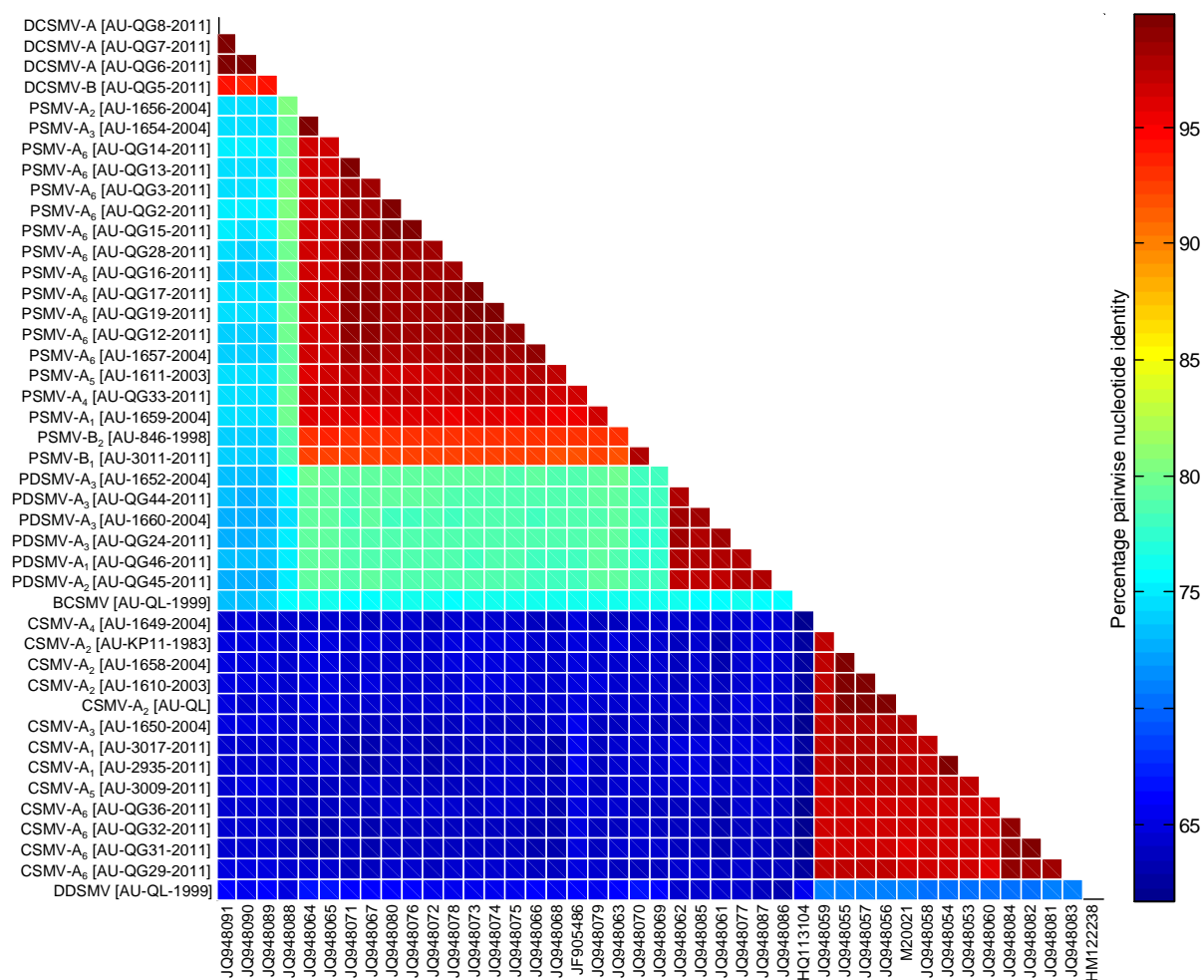


Figure 2.5: Two-dimensional pairwise nucleotide identity plot (percentage identity calculated with pairwise deletion of gaps) comparing degrees of full genome nucleotide sequence similarity amongst Australian monocot-infecting mastrevirus isolates (excluding SSMV-1 and SSMV2).

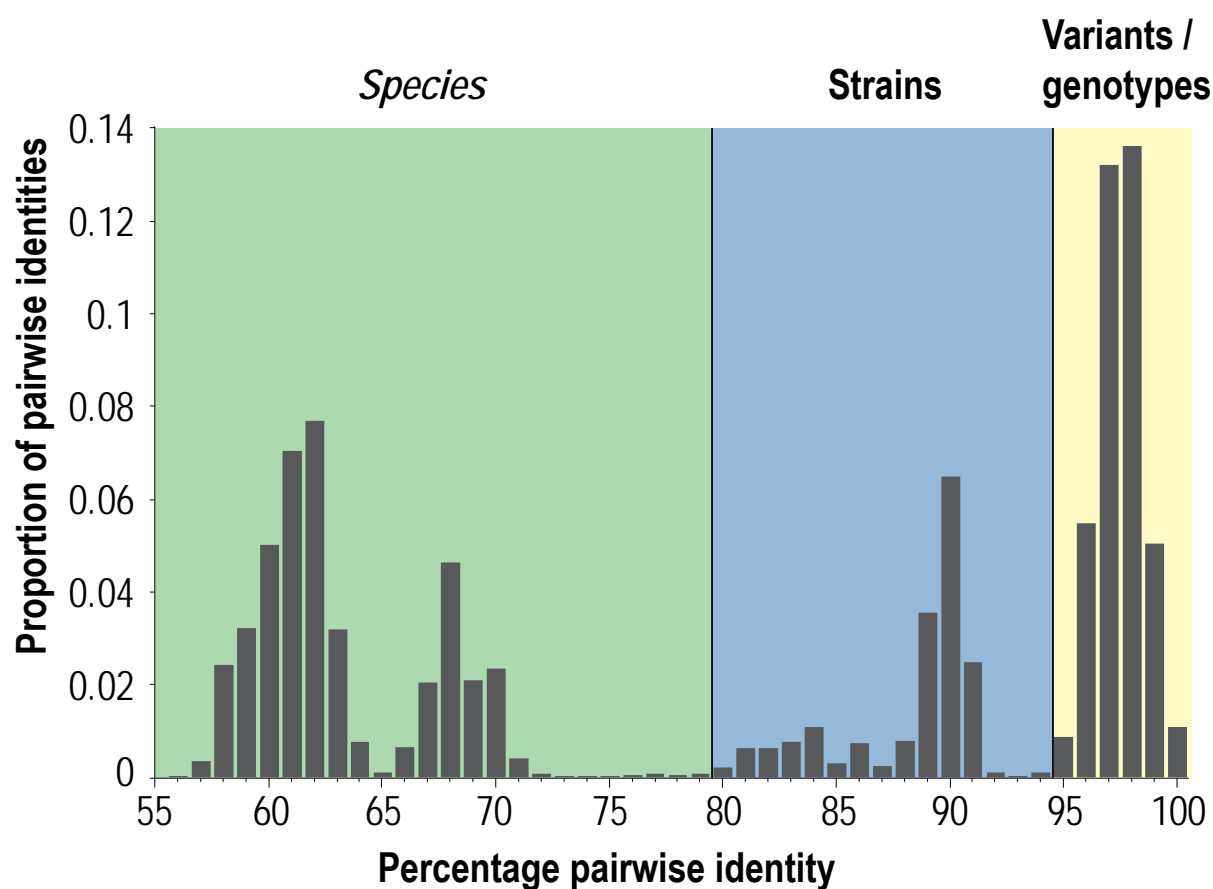


Figure 2.6: Distribution of percentage pairwise nucleotide identities (pairwise deletion of gaps) of all the full genome monocot-infecting mastrevirus sequences (n=568) available in GenBank to determine 161013 pairwise identity comparisons.

2.4.3 SSMV-1 and SSMV-2 resemble divergent African streak viruses

Our phylogenetic analyses indicated that the two highly divergent viruses detected in our survey, SSMV-1 and SSMV-2, are more closely related to African streak viruses than they are to the Australian striate mosaic viruses. The SSMV-1 and SSMV-2 *rep* amino acid sequences are ~54% similar to each other and <53.4 % similar to all other mastrevirus Rep proteins (Fig. 2.7C). On the other hand, the CPs of SSMV-1 and SSMV-2 share 48.6% pairwise amino acid identity to each other and <45% identity to the *cp* of all other mastreviruses (Fig. 2.7C). Using the Rep and full genome sequences, SSMV-1 and SSMV-2 are more closely related phylogenetically to the African streak virus clade (Fig. 2.3A and Fig. 2.7A), whereas using the highly divergent CP sequences, these viruses are sister to both the African streak virus and Australian striate mosaic virus clades (Fig. 2.7B).

Digitaria streak virus from Vanuatu in the South Pacific is most closely related to MSV and conversely, a recently isolated virus from the Caprivi region of Namibia – *Eragrostis minor* streak virus (EMSV) (Martin *et al.*, 2011b) is most closely related to *Miscanthus* streak virus; MiSV) from Japan. This coupled with the fact that the two divergent SSMV isolates are more closely related to African streak viruses than their Australian counterparts, raises an interesting question as to the origin of the African streak viruses. Moreover, based on the diversity of dicot-infecting mastreviruses one could argue that mastreviruses potentially originated in the Australian part of Gondwanaland. However, sampling of mastreviruses in the Middle East and the Indian sub-continent will potentially enable us to generate a more concrete hypothesis on the origin of mastreviruses.

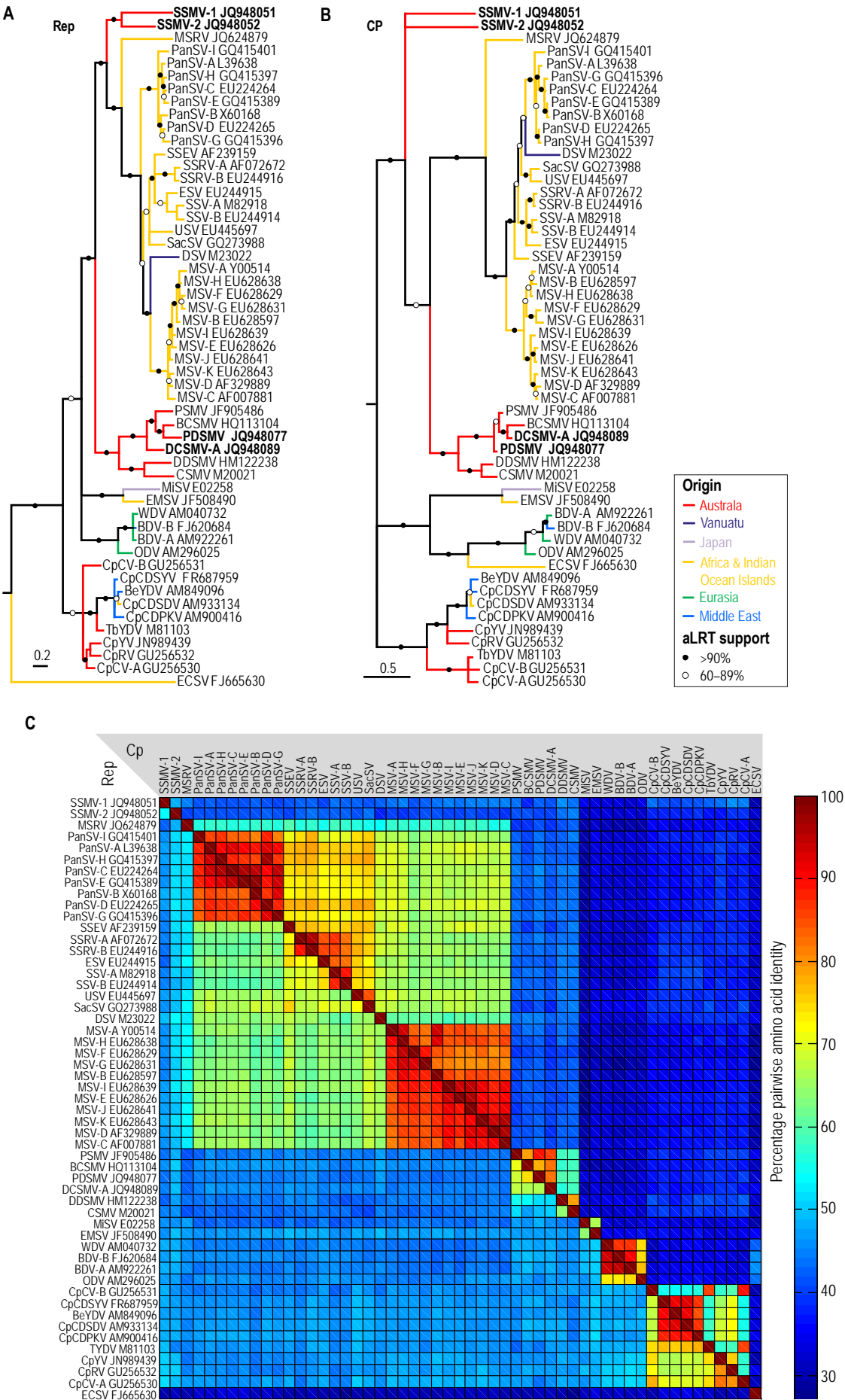


Figure 2.7: See following page for figure legend

Figure 2.7: Maximum likelihood phylogenetic trees of Rep (A) and CP (B) (based on amino acid alignments) depicting the evolutionary relationships of the new Australian monocot-infecting mastrevirus species (represented in bold) to representative mastreviruses from each known species and major strain grouping. Bootstrap support of branches is indicated by open and closed circles, branches with less than 60% bootstrap support have been collapsed. (C) Two-dimensional percentage pairwise amino acid identity plot (with pairwise deletion of gaps) of Rep and CP of the new Australian monocot-infecting mastrevirus species (highlighted in bold) and representative mastrevirus isolates from the known mastrevirus species and major strain groupings.

2.4.4 Evidence of inter- and intra- species recombination

The emergence of new geminiviral strains, and in some cases, species, has largely been attributed to inter- and intra-species recombination (Martin *et al.*, 2011a; Padidam *et al.*, 1999; van der Walt *et al.*, 2009; Varsani *et al.*, 2008b). Therefore we undertook recombination analysis of the 45 available Australian monocot-infecting mastrevirus genomes in order to identify recombination patterns and hotspots.

We detected ten inter-species (Events A-J; Fig. 2.1C; Table 2.2) and nine intra-species (Events C1-2, PD1-3, P1-4; Fig. 2.1C; Table 2.2) recombination events amongst the Australian monocot-infecting mastreviruses. No evidence of recombination was detected in SSMV-1 and SSMV-2. All PSMV isolates have a DCSMV-like region in the C-terminal portion of *repA* (Event J in Fig. 2.1C; Table 2.2). PSMV-A viruses have two regions, one in their *mp* and another in their *cp* that are apparently derived from a PDSMV-like virus (Events F & G in Fig. 2.1C; Table 2.2) whereas PDSMV isolates have two genome regions, one within their *mp* and another within their LIR that are apparently derived from a DCSMV-like virus (Events D & I in Fig. 2.1C; Table 2.2). DCSMV-B is, in turn, apparently a recombinant with a PSMV-like genome region spanning the LIR and the N-terminal encoding portion of *mp* (Event A in Fig. 2.1C; Table 2.2). Intra-species recombination in CSMV and PSMV was observed mainly in the LIR and the C-terminal encoding portions of *repA* and *cp* (Events P1, P2, P3, P4, C1, C2 in Fig. 2.1C; Table 2.2) whereas the three intra-species recombination events detected in PDSMV were distributed throughout the genome (Events PD1-3 in Fig. 2.1C; Table 2.2).

Previous studies have identified found a recombination hotspot at the origin of replication and at the interface between the *cp*/SIR amongst the African streak viruses (Varsani *et al.*, 2009a; Varsani *et al.*, 2009b). The recombination analysis undertaken in this study of the Australian monocot-infecting mastreviruses reveals a clear recombination hotspot in the SIR. Consistent with previous data (Varsani *et al.*, 2009a; Varsani *et al.*, 2008b) we observe that small genomic fragment exchanges (<8% of full genome length) are most common in inter-species recombination events amongst the Australian monocot-infecting mastreviruses, whereas larger fragment exchanges are common in intra-species recombination. The recombination patterns observed amongst the Australian monocot-infecting mastreviruses

closely mirror those found in African streak viruses (Owor *et al.*, 2007a; Shepherd *et al.*, 2008b; Varsani *et al.*, 2009a; Varsani *et al.*, 2008a; Varsani *et al.*, 2008b) and dicot-infecting mastreviruses (Hadfield *et al.*, 2012; Martin *et al.*, 2011a). This suggests that there are selective processes that globally influence mastrevirus recombination patterns which may have been at play since the origin of this genus.

Table 2.2: Details of the recombination events detected using RDP4. Each event is represented by a letter (inter-species) or a letter(s) followed by a number (intra-species). Major and minor parents indicate the approximate identities of parental sequences that respectively donated the larger and smaller fractions of the recombinants genome. The highest p-value indicated for method shown in bold.

Inter-species recombination							
Event	Recombinant (s)	Minor Parent(s)	Major Parent(s)	Breakpoints		Method	P-Value
				Begin	End		
A	DCSMV-B	PSMV	DCSMV-A	2710	458	RGBMCST	4.58X10 ⁻⁸⁸
B	PSMV	EMSV MiSV	PDSMV	1337	1395	RGBMC	6.58X10 ⁻¹⁸
C	PSMV-B	BCSMV	PSMV-A	1396	1453	RGBMCS	3.01X10 ⁻¹⁸
D	PDSMV	DCSMV	PSMV	438	605	RMC	4.53X10 ⁻¹³
E	BCSMV PSMV	DCSMV	PDSMV	1314	1408	RGBM	5.87X10 ⁻¹²
F	PSMV-A	PDSMV	PSMV-B	361	460	RGB	7.26X10 ⁻⁶
G	PSMV-A	PDSMV	PSMV-B	995	1205	RGBM	6.08X10 ⁻⁶
H	BCSMV	PSMV	Unknown	148	466	RBMC	4.52X10 ⁻⁶
I	PDSMV	DCSMV	PSMV	2730	2773	RBMC	1.05X10 ⁻⁶
J	PSMV	DCSMV	PDSMV	2025	2155	RGBMS	6.52X10 ⁻⁵
Intra-species recombination							
CSMV							
C1	CSMV-A ₂ (JQ948055 JQ948056 JQ948057)	CSMV-A ₄	CSMV-A ₂ M20021	1758	1879	GBT	7.76X10 ⁻⁶
C2	CSMV-A ₄	CSMV-A ₆	CSMV-A ₅ , -A ₃	1282	1765	GBMS	3.57X10 ⁻⁵
PDSMV							
PD1	PDSMV-A ₁	PDSMV-A ₃ (JQ948085 JQ948061 JQ948062)	PDSMV-A ₂	2193*	104	RBMS	1.79X10 ⁻⁷
PD2	PDSMV-A ₃ (JQ948077)	PDSMV-A ₂	PDSMV-A ₁ , A ₃ (JQ948062) A ₃ (JQ948085) A ₃ (JQ948087)	837	1687	MCST	3.56X10 ⁻⁵
PD3	PDSMV-A ₃ (JQ948077 JQ948085)	PDSMV-A ₂	PDSMV-A ₃ (JQ948062)	2320	2516	GBMCST	9.15X10 ⁻⁵
PSMV							
P1	PSMV-A ₂ , A ₃ , A ₅ , A ₆	PSMV-B	PSMV-A ₄	1724	1984	RGBMCT	7.45X10 ⁻¹²
P2	PSMV-B ₂	PSMV-A	PSMV-B ₁	753	875	RGBMCST	6.03X10 ⁻⁹
P3	PSMV-A ₅ , A ₆	PSMV-A ₄	PSMV-A ₂ , A ₃ , B ₂	1239*	1476	RGBMCT	6.85X10 ⁻⁶
P4	PSMV-B ₁	PSMV-A ₄	PSMV-B ₂	1766	1809	GBT	7.81X10 ⁻⁵

RDP (R) GENCONV (G), BOOTSCAN (B), MAXCHI (M), CHIMERA (C), SISCAN (S), LARD (L) and 3SEQ (T).

* = The actual breakpoint position is undetermined.

2.4.5 Selection analysis

Comparative selection analyses of the genes from the various Australian monocot-infecting mastrevirus groups displayed similar levels of diversity reported by Hadfield *et al.* (2012) (439 MSV, 39 PanSV, 78 European dwarf virus isolates, 47 dicot-infecting mastreviruses). Our analysis clearly indicates that all the three genes of the Australian-monocot-infecting mastreviruses are evolving under purifying selection ($dN/dS < 1$) (Table 2.3). Grouping PSMV/PDSMV/DCSMV/BCSMV enabled us to compare the selection amongst the Australian monocot-infecting mastreviruses to that undertaken by Hadfield *et al.* (2012). Consistent with observations made by these authors, the *cp* genes are evolving under the highest degree of negative selection in contrast to the *mp* and *rep* genes. Interestingly, the *mp* gene of the PSMV group is evolving under the lowest degree of purifying selection, which is not surprising given that this species has such a broad host range.

Table 2.3: Normalised non-synonymous / synonymous substitution rate ratios within the *cp*, *mp* and *rep* genes of Australian monocot-infecting mastrevirus (PSMV, CSMV and combined dataset of PSMV / PDSMV / DCSMV / BCSMV) compared with other similarly diverse groups of mastrevirus (*Maize streak virus*; MSV, *Panicum streak virus*; PanSV, European dwarf virus; EDV which includes *Wheat dwarf virus*, *Oat dwarf virus* and *Barley dwarf virus*).

Dataset	Gene		
	Movement protein	Coat protein	Replication-associated protein
PSMV	0.417093	0.067579	0.184383
CSMV	0.299978	0.021378	0.150568
PSMV/PDSMV/DCSMV/BCSMV	0.470904	0.121126	0.230959
MSV	0.363271	0.137313	0.174095
PanSV	0.270190	0.117245	0.142445
EDV	0.247922	0.161192	0.188267
Dicot-infecting mastreviruses	0.222441	0.186357	0.223076

virus).

2.4.6 Host range analysis

In the 1988 study undertaken by Greber (1989) investigated the natural host ranges of Australian monocot-infecting mastrevirus (Fig. 2.8A), several classical virology methods were used to verify viral species which included visual symptom assessments, serological assays and/or vector transmission experiments (Greber, 1989; Pinner *et al.*, 1992). Host range studies of five monocot-infecting mastreviruses referred to as CSMV type strain, CSMV *M. stipoides* strain, PSMV, *B. catharticus* geminivirus and *D. didactyla* geminivirus investigating 25 grass species were investigated. The natural host ranges of BCSMV and DDSMV were apparently limited to a single host whereas PSMV and CSMV proved to have a wider host range. Greber (1989) described a CSMV-M variant that was only found in the natural host *Microlaena stipoides* and, based on our natural isolate host range and molecular identity; we have not identified this variant in the wild.

Our surveys revealed that CSMV was predominantly found in *Chloris gayana* (n=7; includes CSMV-A₂ [AU-QLD] M20021) although there was one record from *Digitaria ciliaris*. PSMV was predominantly found in *Paspalum dilatatum* (n=16; includes PSMV-A₅ [AU-1611-2003] JF905486) although there was one record each from *Digitaria ciliaris* and *Ehrharta erecta*. Similar to PSMV, PDSMV was predominantly isolated from *Paspalum dilatatum* (n=5) with one isolate being obtained from *Digitaria ciliaris* (Fig. 2.8B).

Plant species	(A) Greber - Host range study	(B) Natural Host
<i>Aegilops variabilis</i>	→ → → →	
<i>Avena sativa</i>	○ → → → → →	
<i>Brachiaria subquadipara</i>	○ →	
<i>Bromus catharticus</i>	○ × × × × →	①
<i>Chloris gayana</i>	○ ○ → → → → →	⑦
<i>Dactyloctenium australe</i>	○ → × × ×	
<i>Dactyloctenium aegyptium</i>	→ → → → →	
<i>Digitaria ciliaris</i>	× × × × ×	① ① ① ④
<i>Digitaria didactyla</i>	○ × × × × ×	①
<i>Eleusine indica</i>	○ → →	
<i>Ehrharta erecta</i>		①
<i>Eriochloa polystachya</i>		①
<i>Hordeum vulgare</i>	○ → → → →	
<i>Leptochloa filiformis</i>	→ → → → →	
<i>Lolium multiflorum</i>	→ → →	
<i>Microlaena stipoides</i>	○ → × × × ×	
<i>Panicum</i> sp.		①
<i>Panicum miliaceum</i>	→ → →	
<i>Paspalum conjugatum</i>	○ → ×	
<i>Paspalum dilatatum</i>	○ ○ → → × ×	① ①⑥ ⑤
<i>Paspalum longiflorum</i>	○ →	
<i>Paspalum plicatulum</i>	○ →	
<i>Paspalum urvillii</i>	○ →	
<i>Phalaris canariensis</i>	→ → →	
<i>Setaria italica</i>	→ →	
<i>Sporobolus</i> sp.		① ① ①
<i>Triticum aestivum</i>	→ → → → →	①
<i>Urochloa panicoides</i>	→	
<i>Zea mays</i>	○ ○ → → → → →	

○ Natural host	Australian monocot-infecting mastrevirus
→ High transmission efficiency	CSMV BCSMV DCSMV
→ Low transmission efficiency	CSMV-M DDSMV SSMV-1
× Transmission unsuccessful	PSMV PDSMV SSMV-2

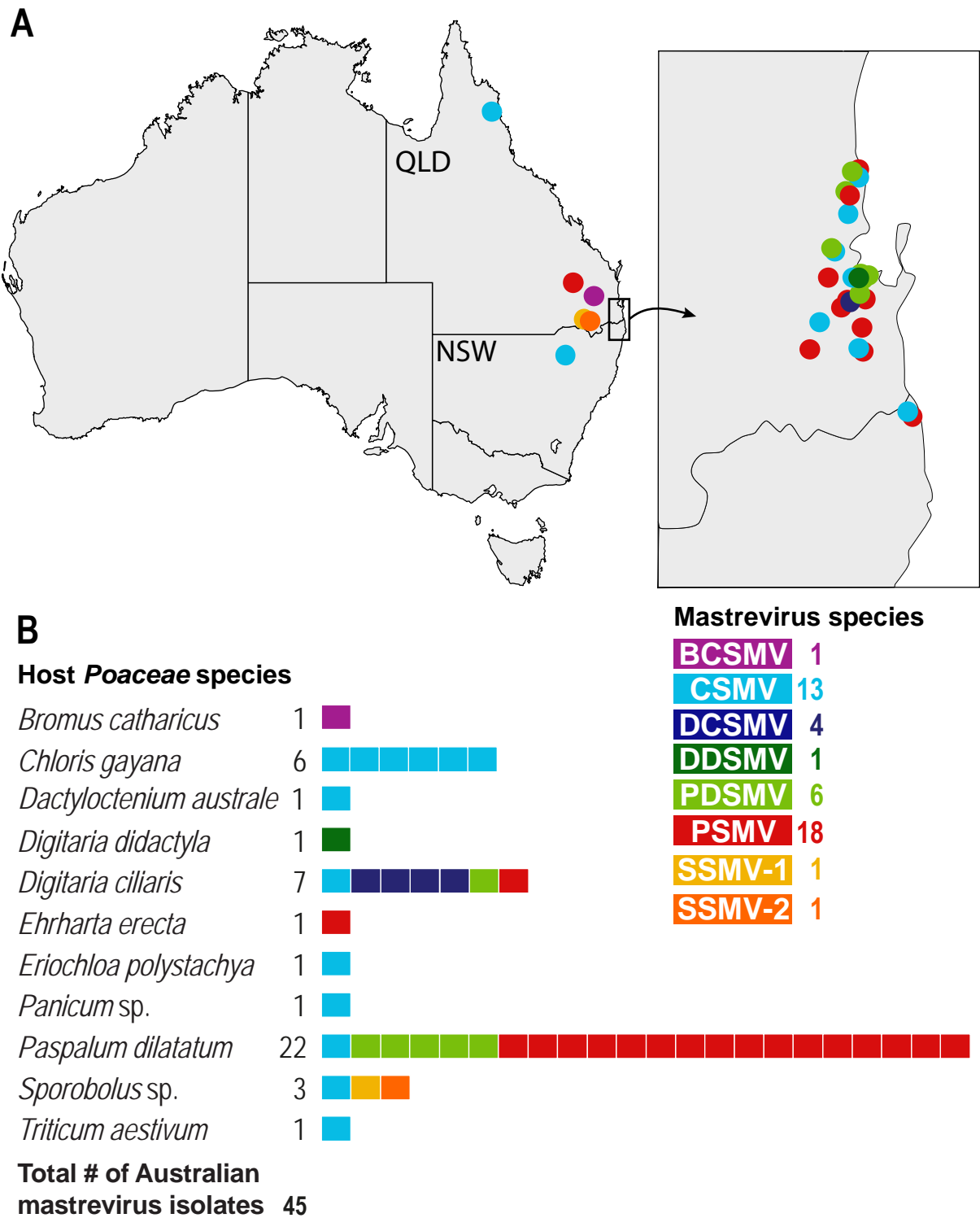
Figure 2.8: Summary of Australian monocot-infecting mastrevirus host range. (A) Previous work undertaken by Greber (1989) showing natural host range, transmission efficiency in different hosts and unsuccessful transmissions for each virus isolate. (B) Natural host range of 41 monocot-infecting mastrevirus isolated from grasses in Australia as part of this study and the four species previously characterised at a molecular level (PSMV, CSMV, BCSMV and DDSMV). Numbers in circles indicate the number of grass samples infected with the different virus species.

2.5 Concluding remarks

Monocot-infecting mastreviruses with broadly overlapping geographical ranges are hosted under natural conditions by a wide variety of Australian grass species. Given that the data presented here was primarily obtained from opportunistic sampling over a relatively small area of eastern Australia (Additional fig. 2.1), the full geographic range of the various virus species described remains to be determined.

Interestingly, Greber (1989) isolated CSMV and PSMV from maize and was able to transmit these as well as BCSMV and DDSMV into maize and wheat. This raises various important questions on the potential for these viruses to become agricultural pest. In Africa, a strain of MSV apparently emerged as a serious maize pathogen following recombination between two grass adapted MSV strains (Harkins *et al.*, 2009a; Varsani *et al.*, 2008b). Despite what must be assumed to have been continual contact between grass adapted MSV strains and maize following the introduction and spread of this species throughout Africa in the 16th and 17th century, the recombination event that finally yielded the maize adapted virus that today threatens African maize production likely only occurred somewhere in southern Africa in approximately the 1850s (Harkins *et al.*, 2009b; Monjane *et al.*, 2011). Since Australia has a shorter history of agriculture than Africa, it is perhaps not surprising that no Australian mastrevirus has yet emerged as a truly economically important agricultural pest. It may therefore be prudent to continually monitor spill-over infections of monocot-infecting mastreviruses in intensively cultivated grasses such as wheat, maize and sugarcane.

Genbank accession numbers: JQ948051 – JQ948091



Additional figure 2.1: (A) Geographical distribution of Australian monocot-infecting mastrevirus isolates in QLD (Queensland) and NSW (New South Wales), Australia. Virus species are represented by different colours. Multiple isolates of the same species from the same location are represented by a single circle. (B) The range of host species from which monocot-infecting mastreviruses were characterised with each coloured blocks representing a single virus isolate of a particular species isolated from that host species.

References

- Accotto, G. P., Donson, J. & Mullineaux, P. M. (1989).** Mapping of Digitaria streak virus transcripts reveals different RNA species from the same transcription unit. *EMBO Journal* **8**, 1033- 1039.
- Andersen, M. T., Richardson, K. A., Harbison, S.-A. & Morris, B. A. M. (1988).** Nucleotide sequence of the geminivirus chloris striate mosaic virus. *Virology* **164**, 443-449.
- Argüello-Astorga, G. R. & Ruiz-Medrano, R. (2001).** An iteron-related domain is associated to Motif 1 in the replication proteins of geminiviruses: Identification of potential interacting amino acid-base pairs by a comparative approach. *Arch Virol* **146**, 1465-1485.
- Boni, M. F., Posada, D. & Feldman, M. W. (2007).** An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176**, 1035-1047.
- Briddon, R. W., Heydarnejad, J., Khosrowfar, F., Massumi, H., Martin, D. P. & Varsani, A. (2010a).** Turnip curly top virus, a highly divergent geminivirus infecting turnip in Iran. *Virus Research* **152**, 169-175.
- Briddon, R. W., Martin, D. P., Owor, B. E., Donaldson, L., Markham, P. G., Greber, R. S. & Varsani, A. (2010b).** A novel species of mastrevirus (family Geminiviridae) isolated from Digitaria didactyla grass from Australia. *Arch Virol* **155**, 1529-1534.
- Brown, J. K., Fauquet, C. M., Briddon, R. W., Zerbini, M., Moriones, E. & Navas-Castillo, J. (2011).** *Virus taxonomy: classification and nomenclature of viruses: Ninth Report of the International Committee on Taxonomy of Viruses*: Elsevier Academic Press, San Diego.
- Dekker, E. L., Woolston, C. J., Xue, Y., Cox, B. & Mullineaux, P. M. (1991).** Transcript mapping reveals different expression strategies for the bicistronic RNAs of the geminivirus wheat dwarf virus. *Nucleic Acids Research* **19**, 4075-4081.
- Edgar, R. C. (2004).** MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797.
- Geering, A. D. W., Thomas, J. E., Holton, T., Hadfield, J. & Varsani, A. (2011).** Paspalum striate mosaic virus: an Australian mastrevirus from Paspalum dilatatum. *Arch Virol* **157**, 193-197.
- Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. (2000).** Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573-582.
- Greber, R. (1989).** Biological characteristics of grass geminiviruses from eastern Australia. *Annals of Applied Biology* **114**, 471-480.
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010).** New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**, 307-321.
- Gutierrez, C. (1999).** Geminivirus DNA replication. *Cellular and Molecular Life Sciences* **56**, 313-329.

- Hadfield, J., Martin, D., Stainton, D., Krabberger, S., Owor, B., Shepherd, D., Lakay, F., Markham, P., Greber, R., Briddon, R. & Varsani, A. (2011).** *Bromus catharticus* striate mosaic virus: a new mastrevirus infecting *Bromus catharticus* from Australia. *Arch Virol* **156**, 335-341.
- Hadfield, J., Thomas, J. E., Schwinghamer, M. W., Krabberger, S., Stainton, D., Dayaram, A., Parry, J. N., Pande, D., Martin, D. P. & Varsani, A. (2012).** Molecular characterisation of dicot-infecting mastreviruses from Australia. *Virus Research* **166**, 13-22.
- Harkins, G. W., Delpont, W., Duffy, S., Wood, N., Monjane, A. L., Owor, B. E., Donaldson, L., Saumtally, S., Triton, G., Briddon, R. W., Shepherd, D. N., Rybicki, E. P., Martin, D. P. & Varsani, A. (2009a).** Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts. *Virology Journal* **6**.
- Harkins, G. W., Martin, D. P., Duffy, S., Monjane, A. L., Shepherd, D. N., Windram, O. P., Owor, B. E., Donaldson, L., van Antwerpen, T., Sayed, R. A., Flett, B., Ramusi, M., Rybicki, E. P., Peterschmitt, M. & Varsani, A. (2009b).** Dating the origins of the maize-adapted strain of maize streak virus, MSV-A. *Journal of General Virology* **90**, 3066-3074.
- Harrison, B. D. (1985).** Advances in Geminivirus Research. *Annual Review of Phytopathology* **23**, 55-82.
- Heyraud, F., Matzeit, V., Schaefer, S., Schell, J. & Gronenborn, B. (1993).** The conserved nonanucleotide motif of the geminivirus stem-loop sequence promotes replicational release of virus molecules from redundant copies. *Biochimie* **75**, 605-615.
- Horváth, G. V., Pettkó-Szandtner, A., Nikovics, K., Bilgin, M., Boulton, M., Davies, J. W., Gutiérrez, C. & Dudits, D. (1998).** Prediction of functional regions of the maize streak virus replication-associated proteins by protein-protein interaction analysis. *Plant Mol Biol* **38**, 699-712.
- Jeske, H., Lütgemeier, M. & Preiß, W. (2001).** DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. *The EMBO journal* **20**, 6158-6167.
- Koonin, E. V. & Ilyina, T. V. (1992).** Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. *The Journal of General Virology* **73** 2763-2766.
- Laufs, J., Traut, W., Heyraud, F., Matzeit, V., Rogers, S. G., Schell, J. & Gronenborn, B. (1995).** In vitro cleavage and joining at the viral origin of replication by the replication initiator protein of tomato yellow leaf curl virus. *Proceedings of the National Academy of Sciences* **92**, 3879-3883.
- Le, S. Q. & Gascuel, O. (2008).** An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution* **25**, 1307-1320.
- Liu, L., van Tonder, T., Pietersen, G., Davies, J. W. & Stanley, J. (1997).** Molecular characterization of a subgroup I geminivirus from a legume in South Africa. *Journal of General Virology* **78**, 2113-2117.
- Martin, D. & Rybicki, E. (2000).** RDP: Detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562-563.

- Martin, D. P., Briddon, R. W. & Varsani, A. (2011a).** Recombination patterns in dicot-infecting mastreviruses mirror those found in monocot-infecting mastreviruses. *Arch Virol* **156**, 1463-1469.
- Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D. & Lefevre, P. (2010).** RDP3: A flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**, 2462-2463.
- Martin, D. P., Linderme, D., Lefevre, P., Shepherd, D. N. & Varsani, A. (2011b).** Eragrostis minor streak virus: an Asian streak virus in Africa. *Arch Virol* **156**, 1299-1303.
- Martin, D. P., Posada, D., Crandall, K. A. & Williamson, C. (2005).** A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Research and Human Retroviruses* **21**, 98-102.
- Martin, D. P., Willment, J. A., Billharz, R., Velders, R., Odhiambo, B., Njuguna, J., James, D. & Rybicki, E. P. (2001).** Sequence diversity and virulence in Zea mays of Maize streak virus isolates. *Virology* **288**, 247-255.
- Monjane, A. L., Harkins, G. W., Martin, D. P., Lemey, P., Lefevre, P., Shepherd, D. N., Oluwafemi, S., Simuyandi, M., Zinga, I., Komba, E. K., Lakoutene, D. P., Mandakombo, N., Mboukoulida, J., Semballa, S., Tagne, A., Tiendrebeogo, F., Erdmann, J. B., van Antwerpen, T., Owor, B. E., Flett, B., Ramusi, M., Windram, O. P., Syed, R., Lett, J. M., Briddon, R. W., Markham, P. G., Rybicki, E. P. & Varsani, A. (2011).** Reconstructing the history of maize streak virus strain a dispersal to reveal diversification hot spots and its origin in southern Africa. *Journal of Virology* **85**, 9623-9636.
- Mullineaux, P. M., Guerineau, F. & Accotto, G.-P. (1990).** Processing of complementary sense RNAs of Digitaria streak virus in its host and in transgenic tobacco. *Nucleic Acids Research* **18**, 7259-7265.
- Mumtaz, H., Kumari, S. G., Mansoor, S., Martin, D. P. & Briddon, R. W. (2011).** Analysis of the sequence of a dicot-infecting mastrevirus (family *Geminiviridae*) originating from Syria. *Virus Genes* **42**, 422-428.
- Nahid, N., Amin, I., Mansoor, S., Rybicki, E., van der Walt, E. & Briddon, R. (2008).** Two dicot-infecting mastreviruses (family *Geminiviridae*) occur in Pakistan. *Arch Virol* **153**, 1441-1451.
- Nash, T. E., Dallas, M. B., Reyes, M. I., Buhrman, G. K., Ascencio-Ibañez, J. T. & Hanley-Bowdoin, L. (2011).** Functional analysis of a novel motif conserved across geminivirus Rep proteins. *Journal of Virology* **85**, 1182-1192.
- Orozco, B. M. & Hanley-Bowdoin, L. (1996).** A DNA structure is required for geminivirus replication origin function. *Journal of Virology* **70**, 148-158.
- Orozco, B. M. & Hanley-Bowdoin, L. (1998).** Conserved Sequence and Structural Motifs Contribute to the DNA Binding and Cleavage Activities of a Geminivirus Replication Protein. *Journal of Biological Chemistry* **273**, 24448-24456.
- Owor, B. E., Martin, D. P., Shepherd, D. N., Edema, R., Monjane, A. L., Rybicki, E. P., Thomson, J. A. & Varsani, A. (2007a).** Genetic analysis of maize streak virus isolates from Uganda reveals widespread distribution of a recombinant variant. *Journal of General Virology* **88**, 3154-3165.

- Owor, B. E., Shepherd, D. N., Taylor, N. J., Edema, R., Monjane, A. L., Thomson, J. A., Martin, D. P. & Varsani, A. (2007b). Successful application of FTA[®] Classic Card technology and use of bacteriophage ϕ 29 DNA polymerase for large-scale field sampling and cloning of complete maize streak virus genomes. *Journal of Virological Methods* **140**, 100-105.
- Padidam, M., Sawyer, S. & Fauquet, C. M. (1999). Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218-225.
- Pinner, M., Markham, P., Rybicki, E. & Greber, R. (1992). Serological relationships of geminivirus isolates from Gramineae in Australia. *Plant Pathology* **41**, 618-625.
- Posada, D. & Crandall, K. A. (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 13757-13762.
- Saunders, K., Lucy, A. & Stanley, J. (1991). DNA forms of the geminivirus African cassava mosaic virus consistent with a rolling circle mechanism of replication. *Nucleic Acids Research* **19**, 2325-2330.
- Schalk, H. J., Matzeit, V., Schiller, B., Schell, J. & Gronenborn, B. (1989). Wheat dwarf virus, a geminivirus of graminaceous plants needs splicing for replication. *EMBO Journal* **8**, 359-364.
- Schwinghamer, M., Thomas, J., Schilg, M., Parry, J., Dann, E., Moore, K. & Kumari, S. (2010). Mastreviruses in chickpea (*Cicer arietinum*) and other dicotyledonous crops and weeds in Queensland and northern New South Wales, Australia. *Australasian Plant Pathology* **39**, 551-561.
- Shepherd, D. N., Martin, D. P., Lefeuvre, P., Monjane, A. L., Owor, B. E., Rybicki, E. P. & Varsani, A. (2008a). A protocol for the rapid isolation of full geminivirus genomes from dried plant tissue. *Journal of Virological Methods* **149**, 97-102.
- Shepherd, D. N., Varsani, A., Windram, O. P., Lefeuvre, P., Monjane, A. L., Owor, B. E. & Martin, D. P. (2008b). Novel sugarcane streak and sugarcane streak Reunion mastreviruses from southern Africa and la Réunion. *Arch Virol* **153**, 605-609.
- Smith, J. M. (1992). Analyzing the mosaic structure of genes. *Journal of Molecular Evolution* **34**, 126-129.
- Stenger, D. C., Revington, G. N., Stevenson, M. C. & Bisaro, D. M. (1991). Replicational release of geminivirus genomes from tandemly repeated copies: evidence for rolling-circle replication of a plant viral DNA. *Proceedings of the National Academy of Sciences* **88**, 8029-8033.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**, 2713-2739.
- Thomas, J., Parry, J., Schwinghamer, M. & Dann, E. (2010). Two novel mastreviruses from chickpea (*Cicer arietinum*) in Australia. *Arch Virol* **155**, 1777-1788.
- van der Walt, E., Rybicki, E. P., Varsani, A., Polston, J. E., Billharz, R., Donaldson, L., Monjane, A. L. & Martin, D. P. (2009). Rapid host adaptation by extensive recombination. *Journal of General Virology* **90**, 734-746.

- Varsani, A., Monjane, A. L., Donaldson, L., Oluwafemi, S., Zinga, I., Komba, E. K., Plakoutene, D., Mandakombo, N., Mboukoulida, J., Semballa, S., Briddon, R. W., Markham, P. G., Lett, J. M., Lefeuvre, P., Rybicki, E. P. & Martin, D. P. (2009a).** Comparative analysis of Panicum streak virus and Maize streak virus diversity, recombination patterns and phylogeography. *Virology Journal* **6**, 194.
- Varsani, A., Oluwafemi, S., Windram, O., Shepherd, D., Monjane, A., Owor, B., Rybicki, E., Lefeuvre, P. & Martin, D. (2008a).** Panicum streak virus diversity is similar to that observed for maize streak virus. *Arch Virol* **153**, 601-604.
- Varsani, A., Shepherd, D. N., Dent, K., Monjane, A. L., Rybicki, E. P. & Martin, D. P. (2009b).** A highly divergent South African geminivirus species illuminates the ancient evolutionary history of this family. *Virology Journal* **6**.
- Varsani, A., Shepherd, D. N., Monjane, A. L., Owor, B. E., Erdmann, J. B., Rybicki, E. P., Peterschmitt, M., Briddon, R. W., Markham, P. G., Oluwafemi, S., Windram, O. P., Lefeuvre, P., Lett, J. M. & Martin, D. P. (2008b).** Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. *Journal of General Virology* **89**, 2063-2074.
- Willment, J. A., Martin, D. P. & Rybicki, E. P. (2001).** Analysis of the diversity of African streak mastreviruses using PCR-generated RFLPs and partial sequence data. *Journal of Virological Methods* **93**, 75-87.
- Wright, E. A., Heckel, T., Groenendijk, J., Davies, J. W. & Boulton, M. I. (1997).** Splicing features in maize streak virus virion- and complementary-sense gene expression. *The Plant Journal* **12**, 1285-1297.
- Xie, O., Suarez-Lopez, P. & Gutierrez, C. (1995).** Identification and analysis of a retinoblastoma binding motif in the replication protein of a plant DNA virus: Requirement for efficient viral DNA replication. *EMBO Journal* **14**, 4073-4082.
- Yazdi, H., Heydarnejad, J. & Massumi, H. (2008).** Genome characterization and genetic diversity of beet curly top Iran virus: a geminivirus with a novel nonanucleotide. *Virus Genes* **36**, 539-545.

Chapter 3

Molecular diversity of monocot-infecting mastreviruses in Africa

Contents

3.1	Abstract.....	115
3.2	Introduction.....	116
3.3	Materials and methods.....	118
3.3.1	DNA extraction and full genome mastrevirus isolation	118
3.3.2	Sequence assembly and phylogenetic analysis.....	119
3.3.3	Host <i>Poaceae</i> species identification	119
3.3.4	Detecting natural selection within the <i>mp</i> , <i>cp</i> and <i>rep</i> codon alignments	119
3.3.5	Recombination analysis.....	120
3.4	Results and discussion	121
3.4.1	Classification of 120 full AfSV genome sequences	121
3.4.2	Diverse host range of MSVs and PanSVs	127
3.4.3	Patterns of geographic distribution.....	129
3.4.4	Conserved patterns of recombination among the monocot-infecting mastreviruses.....	132
3.4.5	Conserved patterns of natural selection signals between MSV and PanSV	140
3.5	Concluding remarks.....	144
3.6	References.....	146

Contributions to this research

I would like to thank the following people for providing me with plant samples which were collected in Africa and the surrounding islands of Gran Canaria, Mauritius and Réunion: D Pande, P Lefeuvre, A Varsani, AL Monjane, S Oluwafemi, S Saumtally, D Linderme, OP Windram and DP Martin.

3.1 Abstract

The most well documented of these viruses is MSV-A which is an important pathogen of maize, a staple crop in Africa. Comparatively not near as much is known about the dynamics and evolution of other monocot-infecting mastreviruses, largely due to the fact that many infect wild uncultivated grasses. In this study we determine the complete sequences of 120 full monocot-infecting mastrevirus genomes from various poaceae species, with the majority from wild uncultivated grasses. This included genomes belonging to the following established species EMSV (n=2), MSRV (n=1), MSV (n=95), PanSV (n=20), SSRV (n=1) and SSV (n=1). We analysed these genomes together with all African monocot-infecting mastreviruses available in GenBank and investigated the geographic distribution, host range and evolutionary dynamics of these viruses. It is evident based on current information that MSV is prevalent and broadly distributed throughout Africa and we now know that its geographic range extends in the north-west to the island of Gran Canaria. Our knowledge regarding the natural host range of both MSV and PanSV has been expanded dramatically including host grasses belonging to an additional 14 genera. Prevalent recombination is apparently occurring among different strains and species and for the first time inter-species recombination events have been detected among species from two geographical locations, Africa and Australia. Indicating ancestors of these viruses once occupied the same region(s) and host(s). Generally MSV and PanSV show similar patterns of natural selection. In the *cp* and *rep* of both species several sites are evolving under episodic diversifying selection pressures which may be indicative of sites which potentially play a role in host specificity. Overall our analyses give an in depth and up to date look at the dynamics and epidemiology of the monocot-infecting mastreviruses and focuses on MSV and PanSV which predominantly infect wild uncultivated grass species.

3.2 Introduction

In Chapter Two monocot-infecting mastrevirus dynamics in Australia was discussed, it was evident in this chapter that Australia is a diversity hotspot for these viruses. It was discussed that Africa is also a known hotspot for monocot-infecting mastrevirus diversity although the major focus of mastreviruses research has gone into the maize pathogen MSV-A. In this study we aim to further elucidate the diversity, host range and evolution dynamics of wild grass adapted monocot-infecting mastreviruses in Africa. The most extensively characterised of the monocot-infecting mastreviruses are the African monocot-infecting mastreviruses, also referred to as the African streak viruses (AfSV). This group is comprised of twelve species which infect various species in the *Poaceae* family. Members of the AfSVs have been recovered from grasses sampled throughout Africa and the surrounding islands (Indian Ocean islands of Réunion and Mayotte) and are vectored by various leafhopper species in the genus *Cicadulina*. The most extensively sample mastrevirus is *Maize Streak virus* (MSV) (Martin *et al.*, 2001; Monjane *et al.*, 2011; Mullineaux *et al.*, 1984; Owor *et al.*, 2007; Varsani *et al.*, 2009; Varsani *et al.*, 2008b) and this can be attributed to MSV's devastating impact on maize production in Africa (Shepherd *et al.*, 2010). It is well documented that only one strain of MSV, MSV-A causes serious disease in maize (Martin *et al.*, 2001), whereas the other identified strains MSV-B – MSV-K cause only mild symptoms in maize and mainly infect uncultivated grass species in the field. With the exception of MSV only one other mastrevirus species *Maize streak Réunion virus* (MSRV) has been isolated from Maize (Oluwafemi *et al.*, 2014; Pande *et al.*, 2012). A significant dataset of *Panicum streak virus* (PanSV) isolates has been collated, consisting of isolates from several countries in Africa and Indian Ocean islands (Varsani *et al.*, 2009; Varsani *et al.*, 2008a). PanSV, as the name suggests has largely been isolated from *Panicum* sp., however, the natural host range of this species extends to several other wild uncultivated grasses (*Brachiaria deflexa*, *Ehrharta calycina*, *Urochloa maxima* and *Urochloa plantaginea*) (Varsani *et al.*, 2009; Varsani *et al.*, 2008a). Of the AfSVs, PanSV has the second highest number of characterised strains, with nine documented strains (PanSV-A through -K).

Other mastrevirus species which have been recovered from wild grasses are *Axonopus compressus streak virus* (ACSV) (Oluwafemi *et al.*, 2014), *Eragrostis minor streak virus*

(EMSV) (Martin *et al.*, 2011b), *Eragrostis streak virus* (ESV) (Shepherd *et al.*, 2008b) and *Urochloa streak virus* (USV) (Oluwafemi *et al.*, 2008).

The remaining five species of the AfSVs, *Sugarcane streak virus* (SSV) (Hughes *et al.*, 1993; Shepherd *et al.*, 2008b), *Sugarcane streak Egypt virus* (SSEV) (Bigarré *et al.*, 1999), *Sugarcane streak Réunion virus* (SSRV) (Bigarré *et al.*, 1999; Shepherd *et al.*, 2008b), *Saccharum streak virus* (SacSV) (Lawry *et al.*, 2009) and *Sugarcane white streak virus* (SWSV) (Candresse *et al.*, 2014), have collectively been referred to as the sugarcane-infecting streak viruses. Although these have predominantly been found infecting sugarcane, SSRV and SSV have been also recovered from wild uncultivated grasses.

Geographically, PanSV and MSV have been identified in several countries throughout Africa and the surrounding islands whereas many of the other species have only been found in a single country. This is most likely due to sampling bias of certain *Poaceae* host species.

Studies have investigated recombination patterns among AfSVs showing that the exchange of genetic fragments between strains and even species is a frequent occurrence and can possibly facilitate the emergence of viruses which are to new host species (Shepherd *et al.*, 2008b; Varsani *et al.*, 2009; Varsani *et al.*, 2008b). A prime example of this is MSV-A which is thought to have emerged as a pathogen of maize following recombination between wild-grass infecting ancestral MSV strains -B and -G/F (Varsani *et al.*, 2008b). An investigation into the historical movement patterns and rates of dispersal of MSV-A subsequent to its emergence has given insight into its epidemiology and highlighted the importance of continued monitoring of these viruses. It may therefore also be equally important to survey AfSVs infecting wild uncultivated grasses to gain an overall picture of the epidemiology of these viruses.

Here we undertook sampling of predominantly symptomatic wild uncultivated *Poaceae* species in the African countries Kenya, Namibia, Nigeria, South Africa, Zimbabwe and the surrounding island nations of Gran Canaria, Mauritius and Réunion, recovering a total of 120 AfSV genomes. Among the 120 genomes recovered we have isolates of EMSV, MRSV,

MSV, PanSV, SSRV and SSV. We analysed these together with all AfSVs available in genbank to identify host ranges, geographic distribution and their evolutionary dynamics. The assemblage of a large MSV and PanSV genomic datasets enabled a robust analysis of natural selection acting on codon sites within the genomes of these viruses. Our investigation extends the known geographic range of several of these species, including the discovery of MSV in Gran Canaria Islands which is the first discovery of an AfSV north-west of the Sahara.

3.3 Materials and methods

3.3.1 DNA extraction and full genome mastrevirus isolation

Poaceae samples displaying foliar striation/streak symptoms that are typical of monocot-infecting mastrevirus infections were collected from Gran Canaria (n=34), Zimbabwe (n=5), Namibia (n=18), Nigeria (n=4), Réunion (n=10), Mauritius (n=40), South Africa (n=95) and Kenya (n=94). Total genomic DNA was extracted from dried leaf material of each sample using either a Extract-N-Amp™ Plant kit (Sigma-Aldrich, USA) as described in Shepherd *et al.* (2008a) or using the GF-1 nucleic acid extraction kit (Vivantis Technologies, Malaysia) according to the manufactures instructions. Circular viral DNA was enriched from the total genomic DNA using the Illustra TempliPhi Amplification Kit (GE Healthcare, USA). Full viral genomes were isolated using either restriction digest or polymerase chain reaction (PCR). For each restriction digestion reaction 1.5µl of templiphi enriched viral DNA was digested using either *Bam*HI, *Kpn*I or *Hind*III to yield unit length ~2.7 kb genomes. PCR was performed using 0.5µl templiphi enriched viral DNA, KAPA HiFi hotstart polymerase (Kapa biosystems, USA) and the degenerate primer pair: dicot forward 5'-GAN TTG GTC CGC AGT GTA GA-3', dicot reverse 5'-GTA CCG GWA AGA CMW CYT GG-3' (Hadfield *et al.*, 2012). The PCR was placed under the following thermocycling conditions: 94°C for 3 min, 25 x [98°C (3 min), 52°C (30 sec), 72°C (2.45 min)] and 72°C for 3 min. Resulting PCR products were purified using the quick-spin PCR Product Purification Kit (iNtRON Biotechnology, Korea) and ligated into pJET1.2 vector (Fermentas, USA). Resulting cloned mastrevirus genomes were Sanger sequencing at Macrogen (Korea) by primer walking.

3.3.2 Sequence assembly and phylogenetic analysis

Full mastrevirus genomes were assembled from overlapping Sanger sequencing reads using DNA Baser sequence assembler V4 (Heracle BioSoft, Romania) and manually managed using MEGA 5.2 (Tamura *et al.*, 2011). Full genome sequence identities were calculated using SDT V1.2 (Muhire *et al.*, 2014). Open reading frames for the movement protein (MP), capsid protein (CP) and replication-associated protein (Rep) were determined using DNAMAN V5.2.9 Lynnon Biosoft, Canada).

A full genome dataset of all monocot-infecting mastreviruses, including those from this study together with all available on Genbank (downloaded 01/08/2014) and that of *Chickpea chlorotic dwarf virus* (KC172668) as an outgroup. These were linearised at beginning of the nonanucleotide sequence (TAATATTAC) and aligned using MUSCLE (Edgar, 2004), implemented in MEGA 5.2 (Tamura *et al.*, 2011). A maximum likelihood (ML) phylogenetic tree was constructed in PhyML version 3.0 (Guindon *et al.*, 2010) using an approximate likelihood ratio test (aLRT) for branch support and the best fit model GTR+G+I591 was chosen by jModelTest (Darriba *et al.*, 2012). Branches with aLRT branch support <80% was collapsed using Mesquite version V1.12.

3.3.3 Host *Poaceae* species identification

Host species were identified for each sample polymerase chain reaction (PCR) amplification of a portion of the chloroplast *ndhF* gene (~1.1kb) was amplified from extractions of total genomic DNA using the primer pair *ndhF*972F and 5'-GTCTCAATTGGGTTATATGAT-3', *ndhF*2110R 5'-CCCCCTAYATATTTGATACCTT-3' (Giussani *et al.*, 2001; Olmstead & Reeves, 1995). 4µl of genomic DNA together with Kapa HiFi hotstart DNA polymerase was put through the following thermocycling conditions: 94 °C for 3 min, 25x (98 °C (20 sec), 50 °C (15 s), 72 °C (1 min)), final extension of 72 °C for 3 min. PCR products were purified using PCR quick-spin Purification Kit (iNtRON Biotechnology Inc, Korea) and sanger sequenced by Macrogen Inc. (Korea).

3.3.4 Detecting natural selection within the *mp*, *cp* and *rep* codon alignments

The full genome sequence datasets of MSV and PanSV were divided into movement protein (*mp*), capsid protein (*cp*) and replication-associated protein (*rep*) coding regions and

realigned based on a codon alignment with MUSCLE (Edgar, 2004). From these alignments the MSV and PanSV datasets were further split for each coding region (*mp*, *cp*, and *rep*) dataset. Recombination breakpoints identified with the GARD method (Kosakovsky Pond *et al.*, 2006) were removed prior to selection analysis. The six datasets were separately analysed for evidence of selection acting on individual codon sites using the MEME (Murrell *et al.*, 2012) and FUBAR (Murrell *et al.*, 2013) methods implemented in the HyPhy package via the online DATAMONKEY server (<http://www.datamonkey.org/>) (Delpont *et al.*, 2010). The FUBAR method was used to identify individual codon sites evolving under either diversifying or negative selection. The MEME method was used to identify individual codons evolving under episodic diversifying selection within individual sub-lineages within the analysed datasets. Due to the low number of genomic sequences available for the species of African monocot-infecting mastrevirus other than MSV and PanSV we were unable to undertake a selection analysis on these species.

3.3.5 Recombination analysis

Full genome datasets of the monocot-infecting mastreviruses were analysed for evidence of recombination events using the platform RDP4 (Martin *et al.*, 2010) with the following methods RDP (Martin & Rybicki, 2000), GENECONV (Padidam *et al.*, 1999), Bootscan (Martin *et al.*, 2005), Maxchi (Smith, 1992), Chimera (Posada & Crandall, 2001), Siscan (Gibbs *et al.*, 2000), LARD (Holmes *et al.*, 1999) and 3Seq (Boni *et al.*, 2007). A dataset comprised of all the monocot-infecting mastreviruses was analysed for evidence of inter-species and intra-species recombination. Additionally a dataset of only MSVs and another only PanSVs was analysed for evidence of intra-species recombination in order to ensure a robust analysis of these two major groupings. Recombination events detected by RDP V4 were deemed credible if the following criteria was met; strong phylogenetic evidence and each event detected by a minimum of three methods with p-values of $<10^{-3}$. Results shown for intra-species recombination excludes events where the MSV-A is the recombinant as extensive MSV-A recombination analyses has previously been undertaken and published (Harkins *et al.*, 2009; Monjane *et al.*, 2011; Varsani *et al.*, 2008b) and this study does not include any additional MSV-A sequences. This is also the case for the Australian monocot-infecting mastreviruses for which intra-species and inter-species (among those found in Australia) has previously been reported by Kraberger *et al.* (2012).

3.4 Results and discussion

3.4.1 Classification of 120 full AfSV genome sequences

A total of 120 full AfSVs genomes were recovered from symptomatic wild uncultivated grasses, maize, sugarcane and one from wheat. Samples from 26 *Poaceae* genera were collected in five Africa countries (Kenya, Namibia, Nigeria, South Africa and Zimbabwe) and three of the surrounding islands (Gran Canaria, Mauritius and Réunion) (Table 3.1). A genome-wide comparison of the 120 recovered genomes with those of previously documented monocot-infecting mastrevirus species was undertaken using SDT V1.2 (Muhire *et al.*, 2014). According to recommendations for mastrevirus classification specified by Muhire *et al.* (2013), the 120 genomes were assigned to the following established species; EMSV (n=2), MSRV (n=1), MSV (n=95), PanSV (n=20), SSRV (n=1) and SSV (n=1). Further, isolates which share >94% identity with previously described strains were assigned the specific strain demarcation. The new MSV isolates from this study were assigned accordingly to the MSV strains -B (n=19), -C (n=34), -D (n=4), -E (n=3), -F (n=30), -G (n=1), -J (n=1) and -K (n=2), and PanSV isolates to strains PanSV-A (n=18), PanSV-C (n=1) and PanSV-H (n=1). The SSRV isolate belongs to SSRV-A, whereas the SSV isolate shares <90% identity with other SSV isolates and therefore we tentatively propose this be assigned to a new strain called SSV-C. This is only the third time that an SSV isolate has been sampled and all three can be assigned to different strains. SSV-A was sampled from a sugarcane sample collected in South Africa and SSV-B from a *Cenchrus myosuroides* sample collected in Réunion. SSV-C, like SSV-A was also recovered from a sugarcane sample collected in South Africa.

These species and strain designations were supported in our phylogenetic analysis. The phylogenetic tree shown in figure 3.1 provides an overview of the phylogenetic relationships between all monocot-infecting mastrevirus. It has previously been highlighted that PanSV diversity is similar to that seen among the MSVs (Fig. 3.2) (Varsani *et al.*, 2009; Varsani *et al.*, 2008a) and it is evident that this is still the case despite the fact that MSV has been the most highly sampled of the monocot-infecting mastreviruses.

Table 3.1: Details of African monocot-infecting mastrevirus isolates recovered in this study.

Species/Strain	Genbank no.	Host	Latitude	Longitude	Sampling year	Country
EMSV	KM230033	<i>Eragrostis</i> sp.	21.286380 S	55.519637 E	2009	Namibia
	KM230032	<i>Eragrostis</i> sp.	7.469199 N	4.556646 E	2009	Namibia
MSRV	KM230031	<i>Digitaria</i> sp.	21.286380 S	55.519637 E	2011	Reunion
MSV-B	KM230030	<i>Digitaria sanguinalis</i>	33.784763 S	20.117002 E	2010	Kenya
	KM229939	<i>Digitaria sanguinalis</i>	33.798896 S	19.874833 E	2009	Kenya
	KM230029	<i>Digitaria tino</i>	33.784763 S	20.117002 E	2008	Mauritius
	KM230028	<i>Digitaria tino</i>	21.251090 S	55.344000 E	2012	Mauritius
	KM230027	<i>Digitaria tino</i>	21.286380 S	55.519637 E	2008	Mauritius
	KM230026	<i>Digitaria tino</i>	26.387340 S	25.019490 E	2008	Mauritius
	KM230025	<i>Digitaria horizontalis</i>	26.387340 S	25.019490 E	2008	Mauritius
	KM230024	<i>Bambusa oldhamii</i>	26.387340 S	25.019490 E	2008	Mauritius
	KM230023	<i>Cenchrus echinatus</i>	27.591950 S	24.752050 E	2008	Mauritius
	KM230020	Unknown host	27.876550 S	24.815740 E	2011	Reunion
	KM230022	<i>Digitaria ciliaris</i>	31.884000 S	18.635900 E	2008	Reunion
	KM230021	<i>Digitaria ciliaris</i>	29.913123 S	31.017825 E	2008	Reunion
	KM230019	<i>Digitaria</i> sp.	17.789790 S	23.343400 E	2012	Reunion
	KM230018	<i>Digitaria</i> sp.	26.387340 S	25.019490 E	2012	Reunion
	KM230017	<i>Ehrharta erecta</i>	27.736090 S	24.784330 E	2009	South Africa
	KM230015	<i>Bromus catharticus</i>	29.065692 S	30.592636 E	2009	South Africa
	KM230014	<i>Chlorocalymma cryptacanthum</i>	29.742150 S	31.035050 E	2009	South Africa
	KM230016	<i>Lolium rigidum</i>	15.434160 S	29.216640 E	2009	South Africa
	KM230013	<i>Digitaria</i> sp.	15.434160 S	29.216640 E	1987	South Africa
MSV-C	KM230012	<i>Digitaria</i> sp.	20.237280 S	57.493920 E	2010	Kenya
	KM229938	<i>Setaria adhaerens</i>	20.237280 S	57.493920 E	2010	Kenya
	KM230010	<i>Zea mays</i>	20.237280 S	57.493920 E	2010	Kenya
	KM230009	<i>Zea mays</i>	20.237280 S	57.493920 E	2010	Kenya
	KM230008	<i>Zea mays</i>	20.237280 S	57.493920 E	2010	Kenya
	KM230007	<i>Zea mays</i>	21.251090 S	55.344000 E	2010	Kenya
	KM230006	<i>Digitaria sanguinalis</i>	20.231554 S	57.506558 E	2010	Kenya
	KM230005	<i>Zea mays</i>	20.231554 S	57.506558 E	2010	Kenya
	KM230004	<i>Zea mays</i>	20.231554 S	57.506558 E	2010	Kenya
	KM230003	<i>Digitaria sanguinalis</i>	20.231554 S	57.506558 E	2010	Kenya
	KM230002	<i>Digitaria didactyla</i>	20.231554 S	57.506558 E	2010	Kenya
	KM230001	<i>Digitaria sanguinalis</i>	20.231554 S	57.506558 E	2010	Kenya
	KM230000	<i>Digitaria</i> sp.	20.231554 S	57.506558 E	2010	Kenya
	KM230011	<i>Zea mays</i>	33.472300 S	20.063000 E	2011	Kenya
	KM229998	<i>Digitaria sanguinalis</i>	33.472300 S	20.063000 E	2009	Kenya
	KM229999	<i>Zea mays</i>	33.472300 S	20.063000 E	2011	Kenya
	KM229997	<i>Zea mays</i>	33.472300 S	20.063000 E	2011	Kenya
	KM229996	<i>Zea mays</i>	33.472300 S	20.063000 E	2011	Kenya
	KM229995	<i>Zea mays</i>	33.472300 S	20.063000 E	2011	Kenya
	KM229937	<i>Digitaria sanguinalis</i>	18.060000 S	20.800000 E	2011	Kenya
	KM229994	<i>Brachiaria deflexa</i>	18.064470 S	21.838550 E	2011	Kenya
	KM229993	<i>Brachiaria deflexa</i>	18.064470 S	21.838550 E	2011	Kenya
	KM229992	<i>Urochloa decumbens</i>	26.051650 S	25.348300 E	2011	Kenya

Table 3.1 continued

Species/Strain	Genbank no.	Host	Latitude	Longitude	Sampling year	Country
MSV-D	KM229991	<i>Hyparrhenia hirta</i>	26.387340 S	25.019480 E	2011	Kenya
	KM229990	<i>Zea mays</i>	26.051650 S	25.348650 E	2011	Kenya
	KM229936	<i>Setaria verticillata</i>	26.051650 S	25.348650 E	2011	Kenya
	KM229989	<i>Setaria barbata</i>	32.180000 S	18.890000 E	2011	Kenya
	KM229988	<i>Urochloa mosambicensis</i>	26.051650 S	25.348650 E	2009	Namibia
	KM229987	<i>Polypogon monspeliensis</i>	31.884000 S	18.635900 E	2009	South Africa
	KM229986	<i>Polypogon monspeliensis</i>	31.884000 S	18.635900 E	2009	South Africa
	KM229985	<i>Polypogon monspeliensis</i>	27.760000 S	30.810000 E	2009	South Africa
	KM229984	<i>Polypogon monspeliensis</i>	26.720000 S	27.100000 E	2009	South Africa
	KM229983	<i>Ehrharta erecta</i>	26.160000 S	27.690000 E	2011	South Africa
	KM229978	<i>Ehrharta erecta</i>	33.784763 S	20.117002 E	2011	South Africa
	KM229982	<i>Polypogon monspeliensis</i>	20.237280 S	57.493920 E	2009	South Africa
	KM229981	<i>Polypogon monspeliensis</i>	20.237280 S	57.493920 E	2009	South Africa
	KM229980	<i>Ehrharta erecta</i>	20.237280 S	57.493920 E	2011	South Africa
	KM229979	<i>Digitaria sanguinalis</i>	20.237280 S	57.493920 E	2011	South Africa
MSV-E	KM229977	Unknown host	20.237280 S	57.493920 E	2011	Reunion
	KM229976	Unknown host	20.237280 S	57.493920 E	2011	Reunion
	KM229975	Unknown host	20.311730 S	57.696250 E	2011	Reunion
MSV-F	KM229974	<i>Digitaria sanguinalis</i>	0.400470 N	36.951290 E	2010	Kenya
	KM229973	<i>Digitaria didactyla</i>	0.229800 N	37.647690 E	2010	Kenya
	KM229972	<i>Digitaria</i> sp.	1.159120 N	36.684160 E	2009	Kenya
	KM229971	<i>Digitaria</i> sp.	0.747220 N	34.163730 E	2009	Kenya
	KM229970	<i>Digitaria</i> sp.	1.229860 N	36.840980 E	2009	Kenya
	KM229969	<i>Digitaria horizontalis</i>	0.209660 N	36.387840 E	2008	Mauritius
	KM229968	<i>Digitaria horizontalis</i>	0.895710 N	37.213310 E	2008	Mauritius
	KM229967	<i>Digitaria horizontalis</i>	0.570040 N	37.188170 E	2008	Mauritius
	KM229966	<i>Digitaria ciliaris</i>	1.054230 N	37.085030 E	2008	Mauritius
	KM229965	<i>Digitaria ciliaris</i>	0.631990 N	37.249840 E	2008	Mauritius
	KM229964	<i>Elusine indica</i>	0.279540 N	36.889190 E	2008	Mauritius
	KM229963	<i>Digitaria horizontalis</i>	0.343760 N	37.629550 E	2008	Mauritius
	KM229961	<i>Digitaria horizontalis</i>	0.445060 N	34.152690 E	2008	Mauritius
	KM229962	<i>Paspalum</i> sp.	3.399830 S	39.495330 E	2012	Mauritius
	KM229960	<i>Digitaria horizontalis</i>	0.106930 N	34.492240 E	2008	Mauritius
	KM229959	<i>Digitaria horizontalis</i>	0.468480 N	34.309670 E	2008	Mauritius
	KM229958	<i>Digitaria horizontalis</i>	20.388650 S	57.584270 E	2008	Mauritius
	KM229957	<i>Digitaria horizontalis</i>	32.366060 S	18.955535 E	2008	Mauritius
	KM229956	<i>Digitaria horizontalis</i>	28.121519 S	15.442484 W	2008	Mauritius
	KM229955	<i>Digitaria horizontalis</i>	20.294170 S	57.532190 E	2008	Mauritius
	KM229954	<i>Digitaria horizontalis</i>	20.294170 S	57.532190 E	2008	Mauritius
	KM229953	<i>Digitaria horizontalis</i>	20.388650 S	57.584270 E	2008	Mauritius
	KM229952	<i>Digitaria horizontalis</i>	20.388650 S	57.584270 E	2008	Mauritius
	KM229951	<i>Digitaria horizontalis</i>	20.388650 S	57.584270 E	2008	Mauritius
	KM229950	<i>Digitaria horizontalis</i>	20.388650 S	57.584270 E	2008	Mauritius
	KM229949	<i>Digitaria horizontalis</i>	20.388650 S	57.584270 E	2008	Mauritius
	KM229948	<i>Digitaria horizontalis</i>	20.388650 S	57.584270 E	2008	Mauritius
	KM229947	<i>Digitaria horizontalis</i>	20.317840 S	57.509680 E	2008	Mauritius
	KM229946	<i>Digitaria horizontalis</i>	20.102880 S	57.586110 E	2008	Mauritius

Table 3.1 continued

Species/Strain	Genbank no.	Host	Latitude	Longitude	Sampling year	Country
MSV-G	KM229945	<i>Digitaria horizontalis</i>	20.115670 S	57.552890 E	2008	Mauritius
	KM229944	<i>Oplismenus burmannii</i>	20.115670 S	57.552890 E	2010	Zimbabwe
	KM229943	<i>Digiteria</i> sp.	0.185000 N	37.958750 E	2008	Gran Canaria
MSV-J	KM229942	<i>Sclerochloa dura</i>	1.046270 N	37.075870 E	2011	Kenya
MSV-K	KM229941	<i>Brachiaria deflexa</i>	4.080880 S	39.374510 E	2011	Kenya
PanSV-A	KM229940	<i>Setaria adhaerens</i>	0.052760 N	37.169950 E	2011	Kenya
	KM229935	<i>Eragrostis minor</i>	21.332376 S	55.470883 E	2009	Namibia
	KM229922	<i>Ehrharta erecta</i>	1.231810 N	34.483020 E	2009	South Africa
	KM229930	<i>Brachiaria deflexa</i>	1.001600 N	34.101010 E	2009	South Africa
	KM229927	<i>Brachiaria deflexa</i>	1.116040 N	34.314500 E	2009	South Africa
	KM229925	<i>Brachiaria deflexa</i>	0.156389 N	34.250830 E	2009	South Africa
	KM229924	<i>Panicum maximum</i>	0.127930 N	35.091450 E	2009	South Africa
	KM229923	<i>Hordeum vulgare</i>	1.040160 N	35.112040 E	2009	South Africa
	KM229921	<i>Brachiaria deflexa</i>	1.259930 N	35.113780 E	2009	South Africa
	KM229920	<i>Brachiaria deflexa</i>	1.036680 N	35.188790 E	2009	South Africa
	KM229919	Unknown host	0.045990 N	35.214360 E	2009	South Africa
	KM229933	<i>Brachiaria deflexa</i>	0.277030 N	36.027940 E	2009	South Africa
	KM229932	<i>Brachiaria deflexa</i>	0.902480 N	35.256370 E	2009	South Africa
	KM229929	<i>Brachiaria deflexa</i>	1.036680 N	35.188790 E	2009	South Africa
	KM229928	<i>Brachiaria deflexa</i>	0.030410 N	36.201370 E	2009	South Africa
	KM229926	<i>Brachiaria deflexa</i>	0.198550 N	35.034970 E	2009	South Africa
	KM229934	<i>Brachiaria deflexa</i>	1.133880 N	35.096180 E	2010	South Africa
	KM229931	<i>Brachiaria deflexa</i>	1.017800 N	35.037600 E	2009	South Africa
	KM229916	<i>Brachiaria deflexa</i>	0.747220 N	34.163730 E	2009	South Africa
PanSV-C	KM229917	<i>Megathyrsus infestus</i>	21.332376 S	55.470883 E	2010	Zimbabwe
PanSV-H	KM229918	<i>Brachiaria deflexa</i>	21.332376 S	55.470883 E	2011	Nigeria
SSRV-A	KM229915	<i>Saccharum hybrid</i>	21.332376 S	55.470883 E	2005	Reunion
SSV-C	KM229914	<i>Saccharum hybrid</i>	21.332376 S	55.470883 E	2008	South Africa

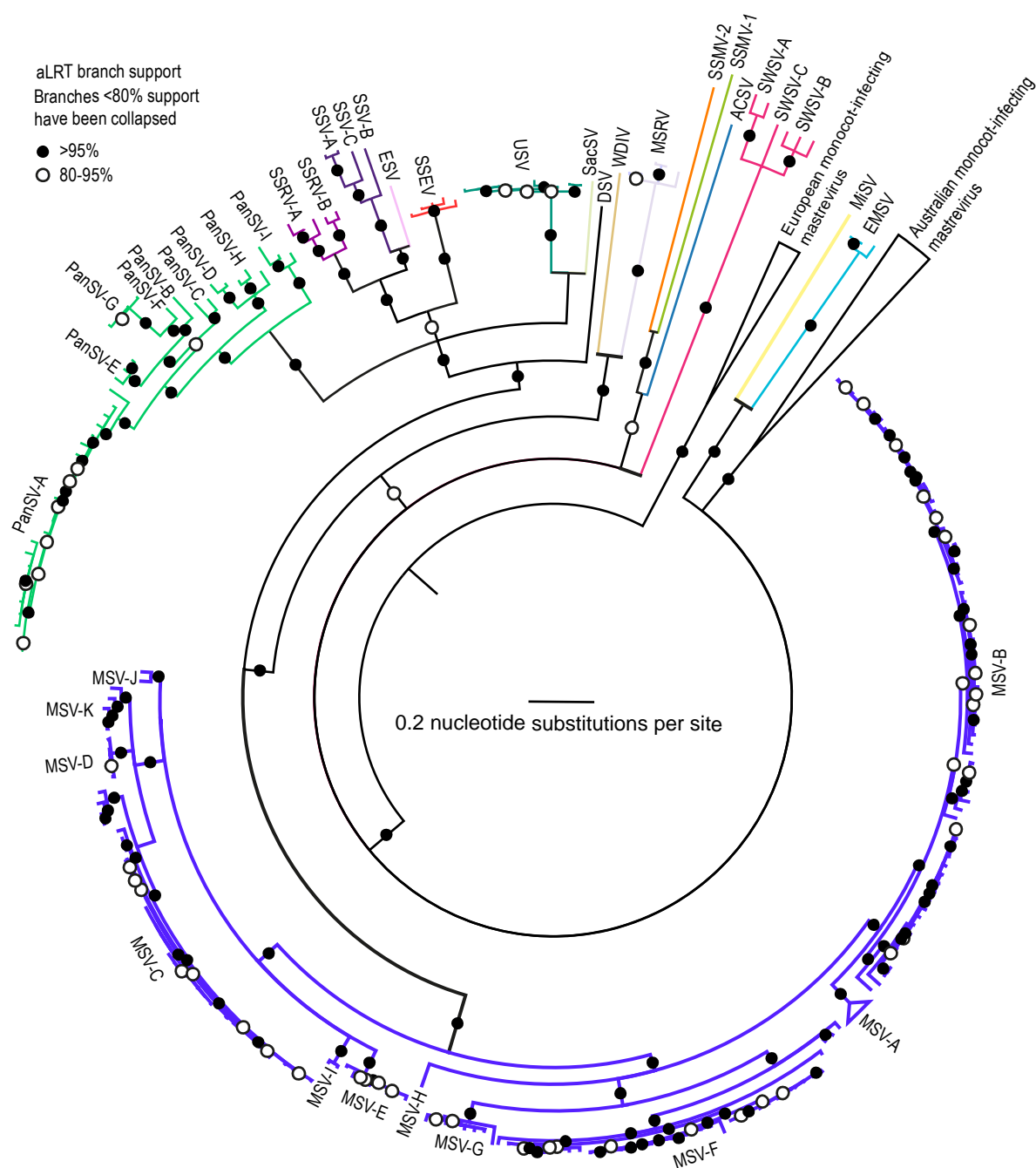


Figure 3.1: Maximum likelihood phylogenetic tree of all available full monocot-infecting mastrevirus genome sequences. Branches are coloured to highlight the different mastrevirus species. All aLRT support branches <80% were collapsed. Australian and European monocot-infecting mastreviruses clades have been collapsed.

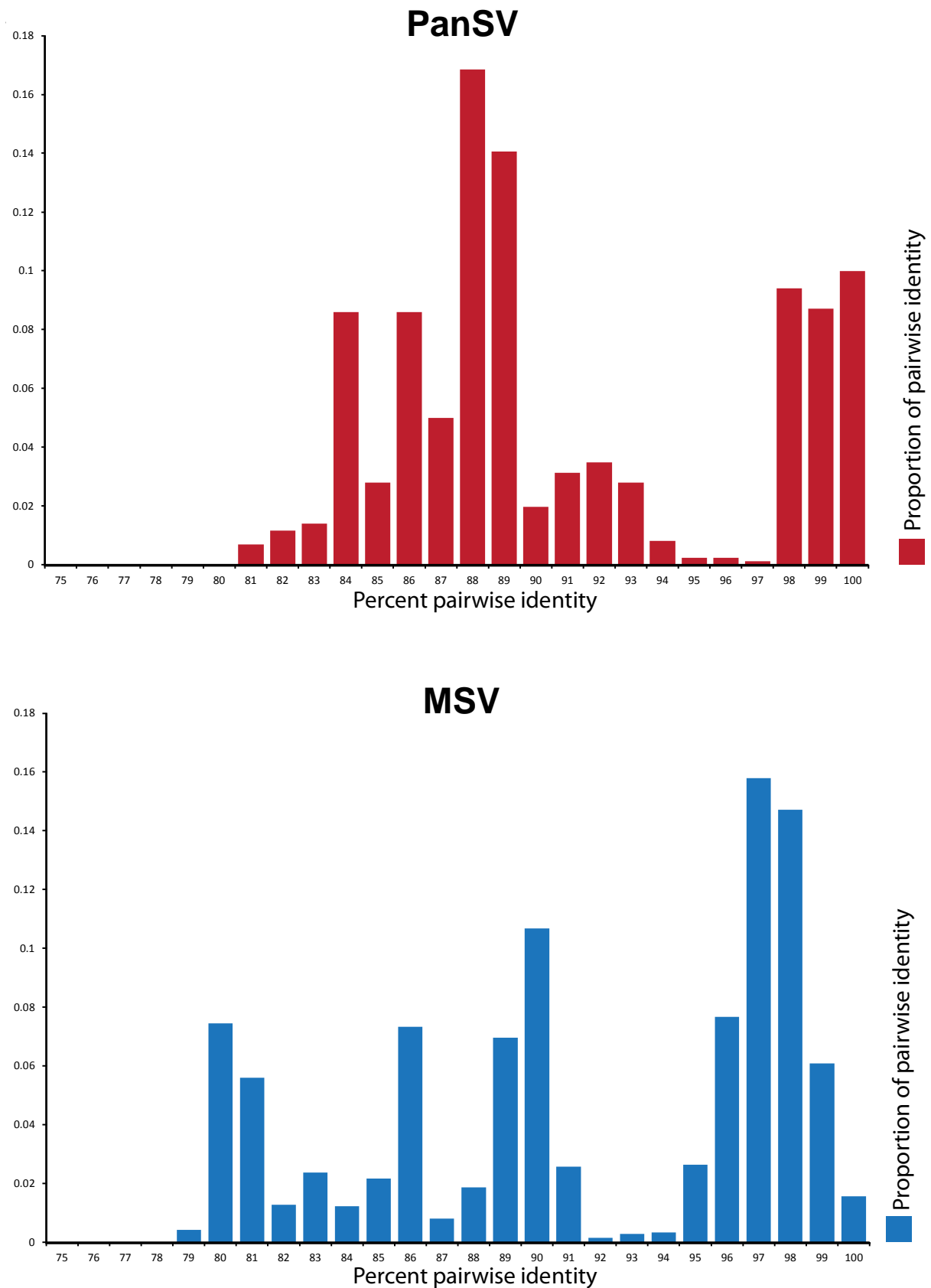


Figure 3.2: Nucleotide pairwise identity distribution plot of full genomes of all maize streak virus and panicum streak virus datasets.

3.4.2 Diverse host range of MSVs and PanSVs

To investigate host range and build on the previous host analysis undertaken by Varsani *et al.* (2008b) we identified the host species (or genus) from which the AfSV genomes were recovered in this study using the chloroplast *ndhF* gene (Table 3.1). Out of the 120 *Poaceae* samples from which mastreviruses were isolated we were unable to identify five host species and hence these are referred to as unknown hosts (Table 3.2). Among the 115 AfSV isolates for which hosts have been identified in this study, there are several newly identified host *Poaceae* species. Twelve isolates of MSV-C were recovered from maize (*Zea mays*). Prior to this study only MSRV, MSV-A, and MSV-B had been isolated from maize. Considering the amount of research undertaken on MSV in maize and the large number of MSV-A isolates recovered from maize (n=330) in the field over the years it is surprising that this strain has not previously been isolated from this host. It may be that other strains such as MSV-C are adapting to new hosts such as maize. Other than maize, three new host species from three different *Poaceae* genera and two subfamilies were identified for MSV-C. This was also the case for a MSV-F found in *Eleusine* sp., *Panicum* sp. and *Paspalum* sp., which had previously only been recovered from *Digitaria* sp. and *Urochloa* sp. It was noted by Varsani *et al.* (2008b) that MSV-A isolates were recovered from grass species spanning eight genera whereas MSV-B was only isolated from grass species spanning six genera, suggesting that MSV-A may have a broader host range. Analysis of a significantly larger dataset which includes all MSV-A (n=377) and B (n=71) isolates recovered to date (available in the GenBank) shows that MSV-A has a host range which includes species from 12 *Poaceae* genera whereas MSV-B has a host range which includes species from 14 *Poaceae* genera. This suggests that MSV-B may in fact have a broader natural host range than MSV-A.

A MSV-B genome was isolated from bamboo (*B. oldhamii*) leaf material collected in Mauritius, to our knowledge this is the first record of any mastrevirus having been isolated from a *Bambusa* sp. Several new host species from three subfamilies were also identified for PanSV-A, with the most common host species identified as *B. deflexa*. Of all the AfSV species, MSV is known to infect grasses from 27 *Poaceae* genera, this is distantly followed by PanSV which is known to infect grasses from nine *Poaceae* genera. It is interesting to note that PanSV and MSV have largely overlapping host ranges, eight of the nine genera PanSV is known to infect are also known to host MSV.

128

Chapter 3

3.4.3 Patterns of geographic distribution

The extensive sampling undertaken in various countries throughout Africa and some of the surrounding Islands in recent years presents a good opportunity to investigate differences in geographic distributions of the AfSVs. In order to gain an overview of this information we mapped all the isolates where full genomes had been recovered of monocot-infecting mastrevirus species in Africa and neighbouring islands (both those recovered in this study and those publically available in GenBank) to indicate country of origin. Additionally we listed the number of species/strains which were recovered in those countries (Fig. 3.3) to get a clear overview. It is worth keeping in mind that some host species have been largely under sampled due to opportunistic sampling biases and others oversampled because of sampling efforts. For example sampling undertaken in Cameroon and Burkina Faso was solely of symptomatic maize material and therefore it is not surprising that only the maize adapted MSV strain (MSV-A) has been found in these countries. Although these sampling bias make it is difficult to compare the diversity identified in one country compared with another, we are able to gain an insight into the distribution of various species and strains.

An extensive analysis of the geographic distribution and the historical movements of MSV-A was undertaken by Monjane *et al.* (2011), their study coupled with that of Harkins *et al.* (2009) determined that MSV-A most likely emerged around ~150 years ago in Southern Africa and subsequently MSV-A spread throughout the continent and the Indian Ocean islands, MSV-A has been sampled in 14 countries in Africa (Monjane *et al.*, 2011; Oluwafemi *et al.*, 2011; Varsani *et al.*, 2009; Varsani *et al.*, 2008b). Wild grass adapted MSV strains MSV-B to -K have also been found to be distributed in various regions of Africa and surrounding islands. Genomes of these MSVs have been sampled in 14 countries including the islands of Mauritius, Réunion and for the first time as part of this study a MSV isolate (MSV-G) was sampled in the island of Gran Canaria. This is the first record of an AfSV found north-west of the Sahara desert. MSV-G has previously only been recorded in Mali and Nigeria (Varsani *et al.*, 2008b) despite the host species (*Digitaria* sp., *Panicum* sp., *Paspalum* sp., and *Brachiaria* sp.) having been sampled multiple times elsewhere in Africa. This strain therefore appears to have a distribution which is restricted to West Africa and the island of Gran Canaria. In this study an additional 37 MSV isolates were recovered from wild grasses and maize and cultivated maize sampled in Kenya. Among these four strains MSV-C, -F, -J

and -K were identified for the first time in Kenya. Additionally, Kenya is the only region where MSV-C has been found infecting Maize.

The islands of the south-west Indian Ocean islands of Africa have previously been shown to be a hotspot of geminivirus diversity (Lefeuvre *et al.*, 2007; Peterschmitt *et al.*, 1996; Shepherd *et al.*, 2008b). Grasses harbouring mastreviruses have been sampled from the islands of Réunion and Mayotte. Réunion has proven to be a hub of diversity with four mastrevirus species identified here (MSRV, MSV, SSRV and SSV). We therefore undertook sampling in the previously unsampled neighbouring island of Mauritius to see what viruses are moving between the islands. Here we recovered MSVs from 32 grass samples spanning five *Poaceae* genera from Mauritius. All of the MSV genomes recovered are MSV-Bs and MSV-Fs.

PanSV was shown as having a similar geographical structure as MSV (Varsani *et al.*, 2009), and it was also noted by these authors that the grass adapted MSV strains are spreading throughout the continent more easily than PanSV. We iterate this observation even with the addition of more PanSV sequences from four countries that the different PanSV strains are apparently regionally constrained.

Africa

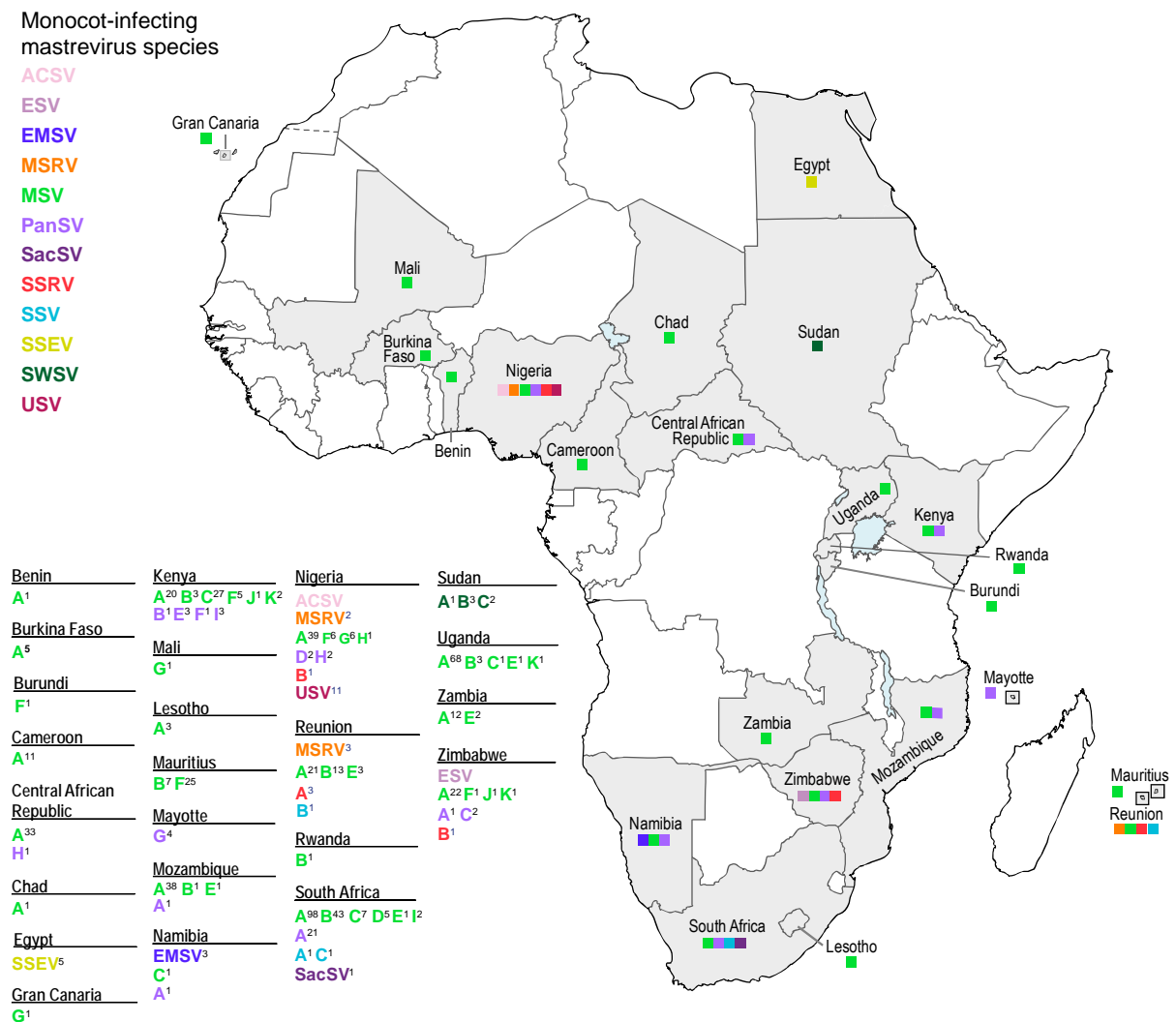


Figure 3.3: Map of Africa and surrounding islands indicating countries where African monocot-infecting mastreviruses have been sampled, this includes samples where full genomes were recovered obtained in this study and those available in the GenBank. Monocot-infecting mastrevirus species are represented by colours in key. Letters in country sample list represent the strain of the virus species of the same corresponding colour shown in key. The number of samples obtained for each species and strain from the various countries is indicated by the superscript number next to virus strain letter or acronym.

3.4.4 Conserved patterns of recombination among the monocot-infecting mastreviruses

Evolution through the mechanism of recombination has been well documented in several species of begomoviruses, curtoviruses and mastreviruses (Kraberger *et al.*, 2013; Lefeuvre *et al.*, 2009; Lefeuvre *et al.*, 2007; Owor *et al.*, 2007; Sanz *et al.*, 2000; Silva *et al.*, 2014; Varsani *et al.*, 2014; Varsani *et al.*, 2009). Among the studies investigating patterns of recombination in mastreviruses are several which have focused on the MSV and PanSV (Monjane *et al.*, 2011; Shepherd *et al.*, 2008b; Varsani *et al.*, 2009; Varsani *et al.*, 2008b). Extensive inter-strain and intra-strain recombination patterns were documented within MSV-A genomes and to a lesser extent those of other MSV strains and PanSV. In this study we aimed to build on previous intra-species recombination analyses of the AfSVs and focus on events occurring among MSV strains other than MSV-A and in PanSV strains and other AfSVs. In addition to this we investigated inter-species recombination events among the monocot-infecting mastrevirus species and for the first time looking at events between species found in two geographically distinct regions of the world, the African originating and the Australian originating monocot-infecting mastreviruses. We therefore analysed the 120 genomes recovered in this study together with the 697 African and Australian monocot-infecting mastrevirus available in GenBank. A total of 47 intra-species (Fig. 3.4; Fig. 3.5; Table 3.3) and 17 inter-species events (Fig. 3.6; Table 3.4) were detected.

Among the intra-species events detected, 23 events were identified in MSV (Fig. 3.4; Table 3.3) (excluding those identified in MSV-A genomes), 16 in PanSV, five in SWSV and one in MSRV, SSV and USV (Fig. 3.5; Table 3.3). This is the first time there has been documented evidence of recombination in MSRV and USV. MSRV and USV isolates share greater than 94% pairwise identity with other members of the species and therefore belong to the same strain, as a result these events would be deemed intra-strain recombination. Only one other intra-strain event was detected, this event involved the exchange of genetic fragments between PanSV-A isolates (event 33 in Fig. 3.5; Table 3.3). It is highly likely that intra-strain recombination occurs more frequently than has documented because it is difficult to detect breakpoints of such exchanges due to the high levels of identity shared among isolates within strains.

Several detected recombinants seem to have resulted from a complex patchwork of intra-species recombination events, for example, some MSV-C recombinants have evidence of two major events which collectively consist of the exchange of genetic fragments which make-up <84 % of the viral genome (Events 3 and 15 in Fig. 3.4; Table 3.3). In PanSV-E recombinant genomes the exchange of genetic fragments overall from two events consists of <64% of the full genome (Events 29 and 26 in Fig. 3.5; Table 3.3).

Varsani *et al.* (2009) described similarities between recombination patterns seen in PanSV and MSV, one key similarity noted was that inter-species events all involved the exchange of genetic material <30% of the full genome. This still remains the case and in fact, it seems to be a pattern which is seen throughout all inter-species recombination events detected in the monocot-infecting mastreviruses (Kraberger *et al.*, 2012; Varsani *et al.*, 2009; Varsani *et al.*, 2008a). Intra-species recombination in monocot-infecting mastreviruses on the other hand tends to involve the exchange of larger fragments which is evident in this analyses, the exchange of genetic material in an event ranges from 1.5% – 50% of the genome. Twelve of the sixteen inter-species events all have breakpoints within the intergenic regions of the genome highlighting a clear hotspot, a common pattern seen in other geminiviruses (Lefeuvre *et al.*, 2009; Martin *et al.*, 2011a; Varsani *et al.*, 2009). A similar hotspot is seen in the intra-species recombination although it is less clear.

It is noteworthy that five of the 16 inter-species events identified here involve the apparent exchange of genetic material between species from two locations which are separated by the Indian Ocean, those found in Africa (including surrounding islands) and those found in Australia (Events E, F, I, Q and R in Fig. 3.6; Table 3.4). Recombination events of a similar nature are also seen among species of dicot-infecting mastreviruses (Kraberger *et al.*, 2013), reinforcing the notion that ancestors of these viral species may at some point have coexisted in the same geographical region(s). Additionally seven inter-species events have inferred parental sequences of monocot-infecting mastreviruses species which are not known to infect the same host *Poaceae* species. For such events to occur the two parental viruses must occupy the same host and therefore this may indicate that some of these species possibly have a broader host range in the field than currently known resulting in an overlap of host species. Ancestors of these inferred parents may also have been able to infect the same host species as seen in events F, I, J, M, O, Q and R in Fig. 3.6 and Table 3.4.

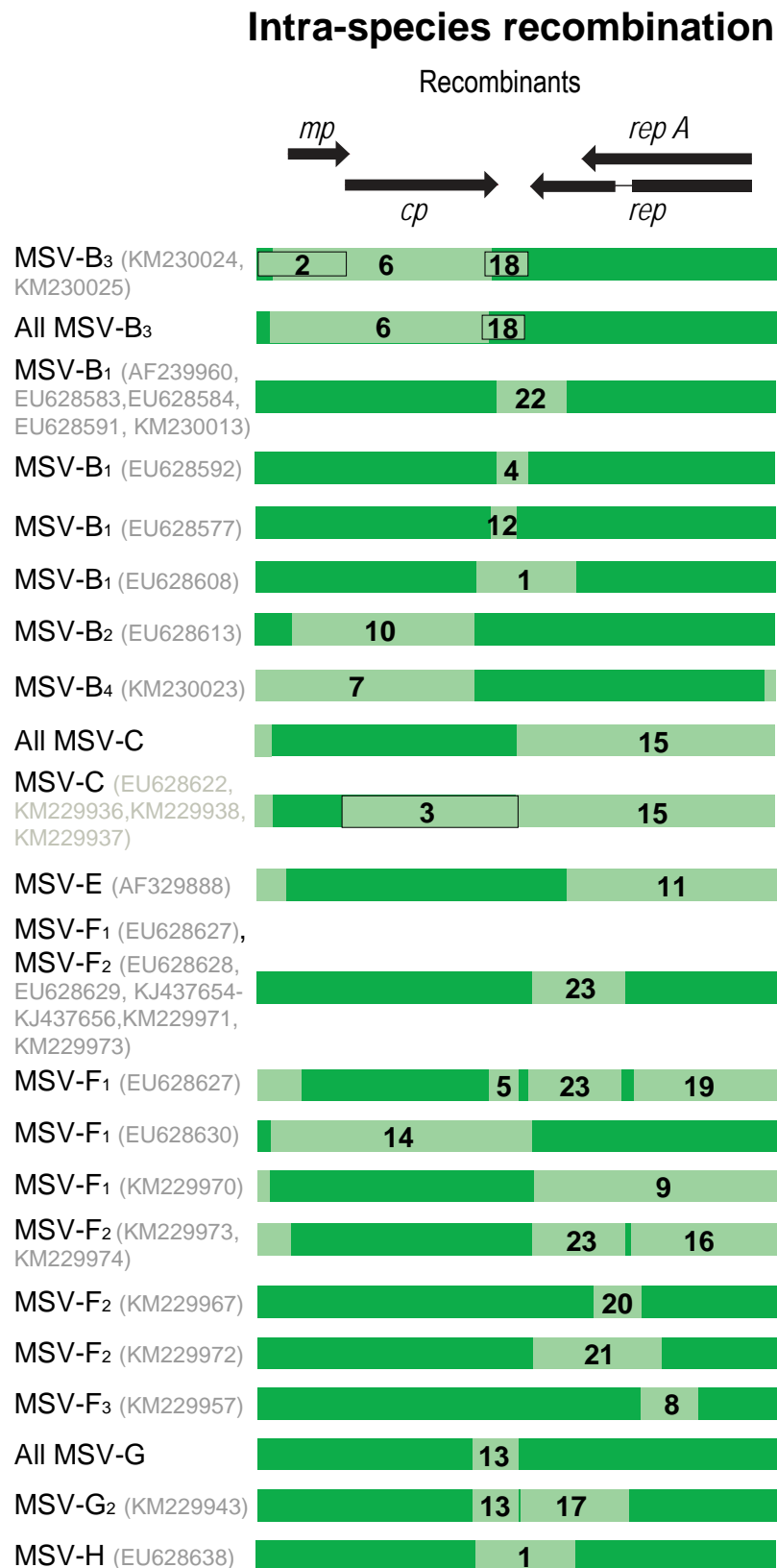


Figure 3.4: Illustration of intra-species recombination events detected among MSV (with the exception of MSV-A). The genome organisation of *mp*, *cp*, *rep* and *repA* in relation to recombinants and recombinant regions is depicted above. Recombinant regions donated by the inferred minor parent is shown in light green, regions donated by the inferred major parent is shown in dark green. Recombination event information for each event can be found in Table 3.3.

Intra-species recombination

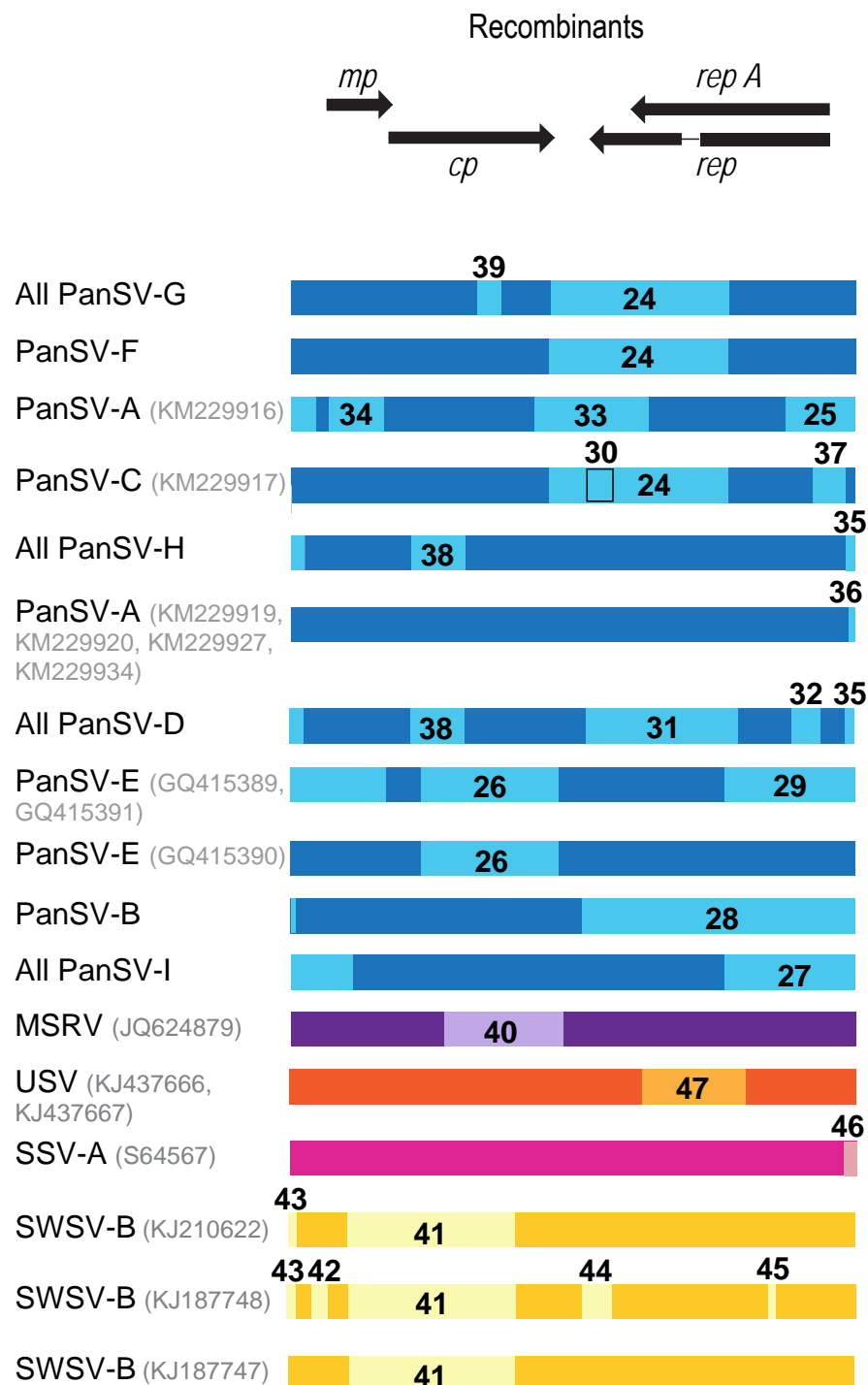


Figure 3.5: Illustration of intra-species recombination events detected among all African monocot-infecting mastrevirus species (with the exception MSV which can be found in Fig. 3.3). The genome organisation of *mp*, *cp*, *rep* and *repA* in relation to recombinants and recombinant regions is depicted above. Recombinant regions donated by the inferred minor parent(s) is shown in lighter shade of colour, region donated by the inferred major parent is shown in darker shade. Each recombination event is indicated by a corresponding number and information for each event can be found in Table 3.3.

Table 3.3: Summary of intra-species recombination events detected among the African monocot-infecting mastreviruses shown in Fig. 3.3 and Fig. 3.4. Major and minor inferred parents represent the likely parents donating the larger and smaller genetic segments of recombinant's genome, respectively. The method with the most significant p-value is indicated in bold and the associated p-value is shown.

Event	Recombinant region	Major Parental Sequence(s)	Minor Parental Sequence(s)	Detection methods	P-value
Intra-species recombination					
1	1235-1743	All MSV-B	All MSV-E, All MSV-J, All MSV-I	RGMCST	1.39×10^{-32}
2	13*-436	All MSV-A, All MSV-B	All MSV-F	RGMST	1.87×10^{-32}
3	456-1372	All MSV-C	All MSV-K	RGMCST	1.38×10^{-14}
4	1288-1429	All MSV-B1	All MSV-A	RGMCST	4.14×10^{-14}
5	1247-1367	All MSV-F	All MSV-A, All MSV-G	RGMCT	3.38×10^{-13}
6	58-1207	All MSV-F4 and MSV-F1	All MSV-A	RGMCST	1.24×10^{-21}
7	2642-1179	All MSV-F1	All MSV-A	RGBMCST	1.16×10^{-22}
8	2025-2329	All MSV-F1, All MSV-F2	All MSV-B1	RGBMCST	3.04×10^{-12}
9	1444*-66	MSV-F2 (EU628628, EU628629, KJ437654, KJ437655, KM229974, KM229973, KM229971, KM229944)	All MSV-B1 (except AF239960, AF239962, EU152260, EU628577, EU628578, EU628582, EU628583, EU628585, EU628591, EU628594, EU628597, EU628598, EU628600, EU628608, EU628609, KM230014, KM230011), MSV-B2 (EU628612)	MCS	1.03×10^{-31}
10	194-1185*	All MSV-F (except KM229974, KM229971, KM229972, KM229944, KM229969, KM229946), All MSV-G	All MSV-A	RGBMCST	1.95×10^{-18}
11	1625-152	All MSV-J	MSV-A6 (AJ225008), MSV-F1 (except KM229971, KM229967, KM229965, KM229963, KM229951, KM229956), MSV-F1 (EU628630, EU628627), All MSV-G	MCS	5.43×10^{-27}
12	1265-1363	MSV-B1 (KM230026, KM230027, KM230029), MSV-B2 (KM230015, AF239960, AF239962, AF329886, EU628578, EU628579, EU628580, EU628581, EU628582, EU628583, EU628584, EU628585, EU628591, EU628594, EU628595, EU628596, EU628597, EU628598, EU628599, EU628600, EU628602)	All MSV-A and All MSV-G	RGBMCT	5.19×10^{-16}
13	1165-1380*	Ancestral MSV-B-like	All MSV-A	RGBMCST	4.06×10^{-14}
14	67*-1438	MSV-B1 (AF239960, AF239962, AF329886, EU628577, EU628578, EU628582, EU628588, EU628597, EU628598, EU628600, EU628609, KM230014, KM230015, KM230016) All MSV-B2	MSV-F2 (EU628628, EU628629, KJ437654, KJ437655, KJ437656, KM229974, KM229973, KM229972, KM229971, KM229969, KM229967, KM229966, KM229965, KM229944, KM229958, KM229959, KM229960, KM229950)	MCS	2.24×10^{-30}
15	1368*-30	All MSV-K, MSV-D (KM229979, KM229980)	Ancestral MSV-A/MSV-B/MSV-G/MSV-F-like	RGBMC	2.78×10^{-10}
16	1979-176	MSV-F1 (EU628627), MSV-F2 (KM229972)	All MSV-G	RMCT	4.48×10^{-09}
17	1374*-1966*	MSV-F1 (EU628627), All MSV-F2 (except KM229964, KM229963, KM229952, KM229945, KM229951, KM229953, KM229961, KM229951, KM229955, KM229954, KM229949, KM229948, KM229956)	Ancestral MSV-B-like	RBCST	2.10×10^{-07}
18	1224-1404	MSV-B1 (AF329887, KM230030, KM230020, KM230013)	MSV-F2 (EU628629, EU628628, KJ437654, KJ437655, KJ437656, KM229972, KM229971, KM229946, KM229944)	RGB	2.04×10^{-15}
19	1993-237	All MSV-F2 (KM229973, KM229974, KM229972, KM229975)	Ancestral MSV-G-like	MCT	7.54×10^{-08}
20	1771*-2035	All MSV-G	MSV-B2 (EU628613, EU628611)	RGBMST	4.81×10^{-05}
21	1439-2141	All MSV-F2 (KM229973, KM229974, KM229972, KM229975)	MSV-B1 (KM230024, KM230025) All MSV-B2	RMC	1.13×10^{-06}
22	1288-1577*	All MSV-B2	MSV-F1 (EU628630), MSV-F2 (EU628629, KJ437655, KJ437656, KM229973, KM229972, KM229971, KM229970, KM229967, KM229965, KM229964, KM229962, KM229958, KM229956, KM229947, KM229946, KM229944), MSV-F3	RGBM	2.94×10^{-05}

Table 3.3 continued

23	1457-1707*	MSV-F2 (KM229965-KM229969, KM229944, KM229946, KM229950, KM229958, KM229959, KM229962, KM229960)	All MSV-B2	RBM	6.51x10 ⁻⁰⁵
24	1333-1969	PanSV-C (EU224264)	Ancestral PanSV-H-like	RGMCST	1.13x10 ⁻¹⁷
25	2383-121	All PanSV-A (except KM229916)	Ancestral PanSV-H-like	RGBMCST	4.24x10 ⁻¹⁹
26	629-1294	Ancestral PanSV-A-like	PanSV-B	RGMCST	1.04x10 ⁻¹⁶
27	2094-190	PanSV-E (GQ415390)	Ancestral PanSV-B/PanSV-C/PanSV-F/PanSV-G-like	RGMCS	5.96x10 ⁻²¹
28	1397-28	PanSV-F	Ancestral PanSV-F	RMCST	6.06x10 ⁻¹³
29	2089-469	All PanSV-I	PanSV-F (GQ415392), PanSV-G (GQ415393, GQ415395)	RGMCS	2.36x10 ⁻²⁰
30	1419-1542	All PanSV-A	Ancestral PanSV-H/PanSV-E-like	RGMC	5.66x10 ⁻⁰⁸
31	1389-2168	All PanSV-H	All PanSV-G	RMCST	2.26x10 ⁻¹⁶
32	2424-2563	Ancestral PanSV-E-like	PanSV-C (KM229917)	RGMCST	2.25x10 ⁻⁰⁴
33	1183-1718	Ancestral PanSV-A-like	PanSV-A (L39638)	GBT	1.38x10 ⁻⁰⁵
34	176*-456	PanSV-A (GQ415386, GQ415387, KM229926, KM229928, KM229929, EU224263)	PanSV-F (GQ415392)	RBL	1.93x10 ⁻⁰⁴
35	2668*-67	PanSV-B (X60168)	PanSV-A (KM229916)	MLT	1.27x10 ⁻⁰⁴
36	2675-10	PanSV-A (GQ415386 L39638)	Ancestral PanSV-A-like	GLT	1.13x10 ⁻⁰⁶
37	2526-2663	PanSV-C (EU224264)	Ancestral PanSV-A-like	GBMC	8.64x10 ⁻⁰⁶
38	576-848	All PanSV-I	PanSV-B, PanSV-F, PanSV-G (GQ415394)	RBMCS	1.83x10 ⁻⁰⁸
39	905-1022*	PanSV-B, PanSV-E (GQ415389, GQ415390), PanSV-F	Ancestral PanSV-A-like	RMCT	1.32x10 ⁻⁰³
40	608-1322*	MSRV (KJ437670, KJ437669)	MSRV (JQ624880, KM230031)	RGMCST	5.50x10 ⁻¹¹
41	278-1080*	SWSV-C (KJ187749)	Ancestral SWSV-A-like	RGBS	4.01x10 ⁻⁰⁶
42	112-178	SWSV-B (KJ210622, KJ187747)	Ancestral SWSV-A-like	RGM	4.06x10 ⁻⁰⁶
43	2827-34	SWSV-B (KJ187747)	SWSV-A (KJ187746, KJ187745)	RGM	2.26x10 ⁻⁰⁸
44	1420*-1540	SWSV-B (KJ210622, KJ187747)	Ancestral SWSV-A-like	RGM	9.54x10 ⁻⁰⁶
45	2505-2559	SWSV-B (KJ210622, KJ187747)	Ancestral SWSV-A-like	RGB	1.70x10 ⁻⁰⁸
46	2698-2752	SSV-A (M82918)	Ancestral SSV-C-like	RGBMCST	3.18x10 ⁻¹⁰
47	2023*-2427	USV (EU445692)	USV (EU445693-EU445699, KJ437665)	RGBMC	2.82x10 ⁻⁰⁸

RDP (R) GENCONV (G), BOOTSCAN (B), MAXCHI (M), CHIMERA (C), SISCAN (S), LARD (L) and 3SEQ (T).

* = The actual breakpoint position is undetermined.

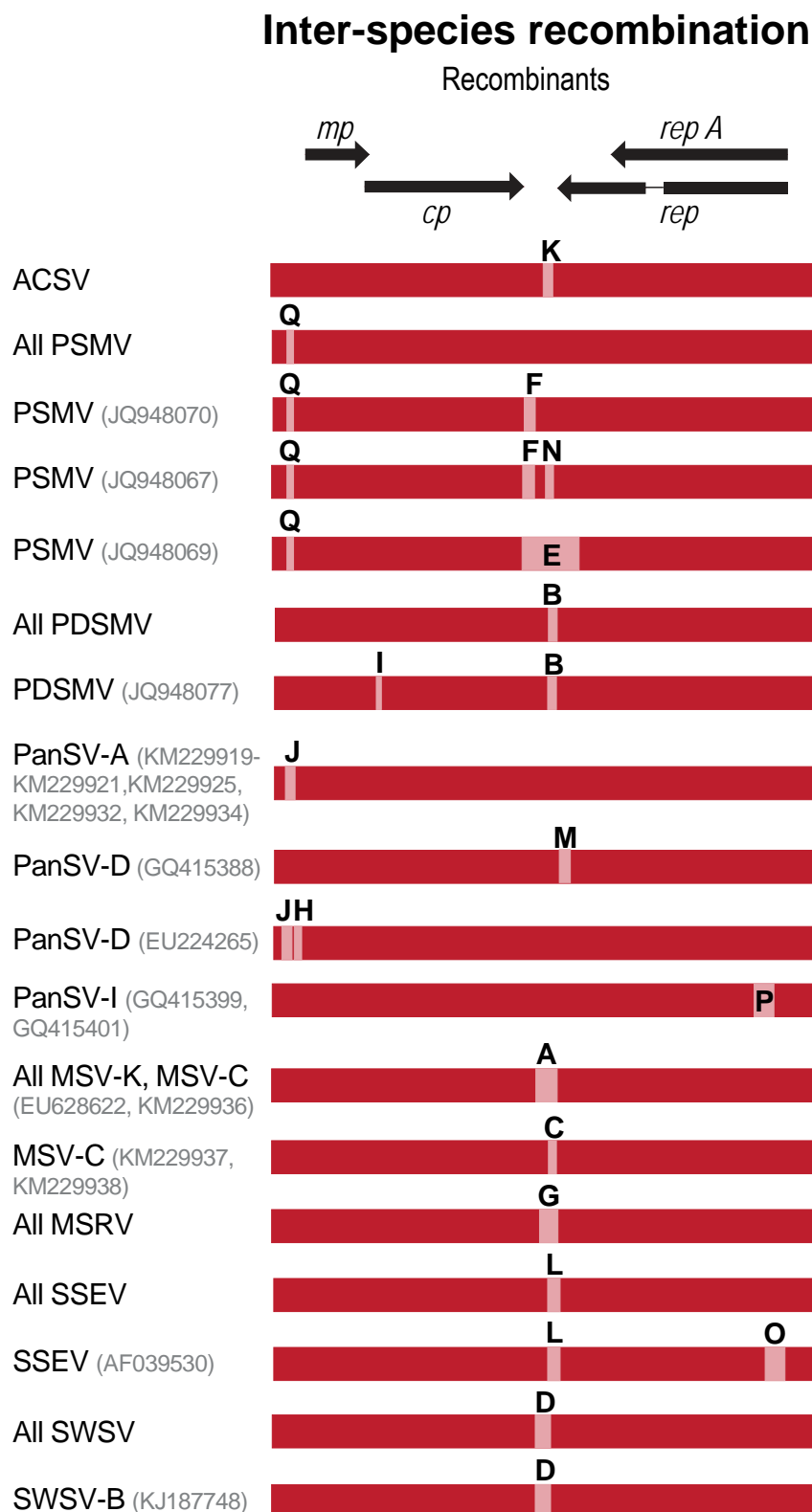


Figure 3.6: Illustration of inter-species recombination events detected among all African and Australian monocot-infecting mastrevirus species (excludes those events detected among the Australian monocot-infecting mastreviruses which has previously been documented by Kraberger *et al.* (2012)). The genome organisation of *mp*, *cp*, *rep* and *repA* in relation to recombinants and recombinant regions is depicted above. Recombinant regions donated by the inferred minor parent is shown in light red, regions donated by the inferred major parent(s) is shown in dark red. Recombination event information for each event can be found in Table 3.4.

Table 3.4: Summary of inter-species recombination events detected among the African and Australia monocot-infecting mastrevirus shown in Fig. 3.5 (excludes those events detected among the Australian monocot-infecting mastreviruses which has previously been documented by Krabberger *et al.* (2012). Major and minor inferred parents represent the likely parents donating the larger and smaller genetic segments of recombinant's genome, respectively. The method with the most significant p-value is indicated in bold and the associated p-value is shown.

Event	Recombination region	Major Parental Sequence(s)	Minor Parental Sequence(s)	Detection methods	P-value
Inter-species recombination					
A	1263-1345	SSV-A, All MSV-A1, A2, A3 and A4, All MSV-D, All MSV-B1, All MSV-E, MSV-H, All MSV-I, All MSV-J, MSV-C (AF007881, KM229983-KM230012)	Ancestral MSRV-like	RGBM	1.27×10^{-18}
B	1320*-1378	Ancestral SSV-A-like, MSV-like	All MSV-F and MSV-B3	RGBMCT	9.83×10^{-18}
C	1300-1341	All MSV-D	ACSV (KJ437671)	RGM	3.86×10^{-08}
D	1355-1419	Ancestral MSV-C-like	All MSV, All SSV-A	RGBM	4.65×10^{-13}
E	1233-1465	All DCSMV-A	Ancestral SSRV/MSV/USV/PanSV/ESV-like	RGB	1.30×10^{-06}
F	1247-1298	PSMV (JF905486, JQ948063-JQ948080)	All USV, PanSV-A (EU224263, GQ415387)	RGM	5.08×10^{-12}
G	1337-1426	Ancestral MSV-C-Like	All MSV-C and MSV-D	RGBMS	1.21×10^{-11}
H	71-99	PanSV-D (GQ415388), PanSV-H (GQ415397)	All MSV, SSV-A	RGBS	2.33×10^{-06}
I	432-465	PDSMV (JQ948061, JQ948062, JQ948085- JQ948087)	ESV (EU244915), SacSV (GQ273988), PanSV-H (GQ415397)	RGM	1.98×10^{-10}
J	43-97	PanSV-A (L39638, GQ415387, EU224261, EU224263, KM229929)	SSRV-B (KJ437668)	RGBM	1.99×10^{-11}
K	1386-1431	Ancestral MSV-C/MSV-B-like	All MSV-C, MSV-D, MSV-D, All MSV-J, All MSV-I	RGB	2.12×10^{-09}
L	1322-1368	All MSV-A, MSV-B (EU628613, AF329887, EU628608, EU628610-EU628612, KM230020, KM230030), All MSV-E, All MSV-G, All MSV-H, All MSV-I, All MSV-J	All PSMV	RGB	2.50×10^{-09}
M	1367-1412	PanSV-D (EU224265)	Ancestral SacSV-like	RGBMCS	2.92×10^{-09}
N	1347-1384	All DCSMV-A	All EMSV	RGBM	1.75×10^{-07}
O	2541-2587	SSEV (AF037752, AF039528, AF039529 AF239159)	PanSV-A (KM229926, KM229928, KM229929), PanSV-H (GQ415397), PanSV-D (GQ415388)	RGB	4.66×10^{-07}
P	2533-2573	PanSV-I GQ415400	Ancestral All MSV/SSV-A-like	RGB	1.28×10^{-06}
Q	69- 104*	PSMV (JQ948069)	All MSRV	RGMC	3.12×10^{-07}

RDP (R) GENCONV (G), BOOTSCAN (B), MAXCHI (M), CHIMERA (C), SISCAN (S), LARD (L) and 3SEQ (T).

* = The actual breakpoint position is undetermined.

3.4.5 Conserved patterns of natural selection signals between MSV and PanSV

In addition to recombination, genetic drift coupled with natural selection is an important mechanism of geminivirus evolution (Duffy & Holmes, 2009; Duffy *et al.*, 2008; Lima *et al.*, 2013). We therefore investigated the selection pressures acting on codon sites within the encoding regions of monocot-infecting mastreviruses by undertaking a comparative selection analysis of the *mp*, *cp* and *rep* of MSV and PanSV (Fig. 3.7). We used the selection detection models FUBAR (Murrell *et al.*, 2013) and MEME (Murrell *et al.*, 2012) to identify signals of selection influencing individual codon sites within the genes of MSV and PanSV. For both the MSV and PanSV we have large enough full genome datasets in order to detect significant signals of selection across a large proportion of the codon sites within the genes of MSV and PanSV.

Selection analysis previously undertaken on mastrevirus species has looked at the overall selection pressure acting on the *mp*, *cp* and *rep* (Hadfield *et al.*, 2012; Kraberger *et al.*, 2012). Results showed that all genes have normalised non-synonymous/synonymous (dN/dS) values less than one which indicates these genes are evolving predominantly under negative selection, also referred to as purifying selection. This has also been documented in other ssDNA viruses (Duffy *et al.*, 2008; Shackelton *et al.*, 2005; Stenzel *et al.*, 2014). Our results show that the *cp* is evolving under the highest degree of purifying selection (average dN/dS of MSV=0.24 and PanSV =0.26), followed by the *rep* (Average dN/dS of MSV=0.36 and PanSV =0.29) and *mp* (Average dN/dS of MSV=0.68 and PanSV =0.54). The *mp* seems to be evolving under the least amount of purifying selection of all three genes in all mastrevirus species investigated (Hadfield *et al.*, 2012; Kraberger *et al.*, 2012).

Codon sites evolving under negative selection seem to be fairly evenly distributed throughout the *cp* and *rep*, with a high proportion of sites (Fig. 3.7; shown in red) under negative selection for the same amino acid in both MSV and PanSV. A notable region which has a high proportion of codon sites evolving under negative selection for the same amino acid in both MSV and PanSV is in the *cp* between codon site 45 and 190. This region spans the majority of the β -barrel structure which is integral to the core structure of the viral capsid (Bennett *et al.*, 2008; Zhang *et al.*, 2001) and therefore preservation of certain amino acids within this motif may be essential for maintaining CP structural integrity in these monocot-

infecting mastrevirus species. Within the Rep many of the functional motifs appear to be predominantly undergoing negative selection (Fig. 3.7; shown in red, orange and grey). Motif C and the region flanking it has a high concentration of codons where strong signals of negative selection were detected (codon 195–codon 320), this clustering in PanSV is in the region of overlap between the *rep* and *repA* and therefore could be an artefact of this overlap (Fig. 3.7; indicated by a grey shaded area), however in MSV there is no overlap between the *rep* and *repA* and therefore this may represent a larger conserved functional region.

Several sites within each gene are also evolving under detectably positive selection, either positive also referred to as diversifying (Fig. 3.7; shown in blue) or episodic diversifying (shown in green). Both favour change in residues at a site (sites where $dN/dS > 1$), however, episodic diversifying selection is acting on a site only in specific subset(s) of the population. Very few sites have been detected to be undergoing positive selection (PanSV: $mp=1$, $cp=1$, $rep=3$; MSV: $mp=3$, $cp=2$, $rep=4$) whereas a significant portion of sites were detected to be evolving under episodic diversifying (PanSV: $mp=6$, $cp=12$, $rep=18$, MSV: $mp=7$, $cp=11$, $rep=23$). Strains of both PanSV and MSV are known to have different host ranges and preferences, it is therefore not surprising that some sites are in a state of change in different subsets, possibly strain subset, in these species and may mean that these sites are varied to be optimal in the different hosts.

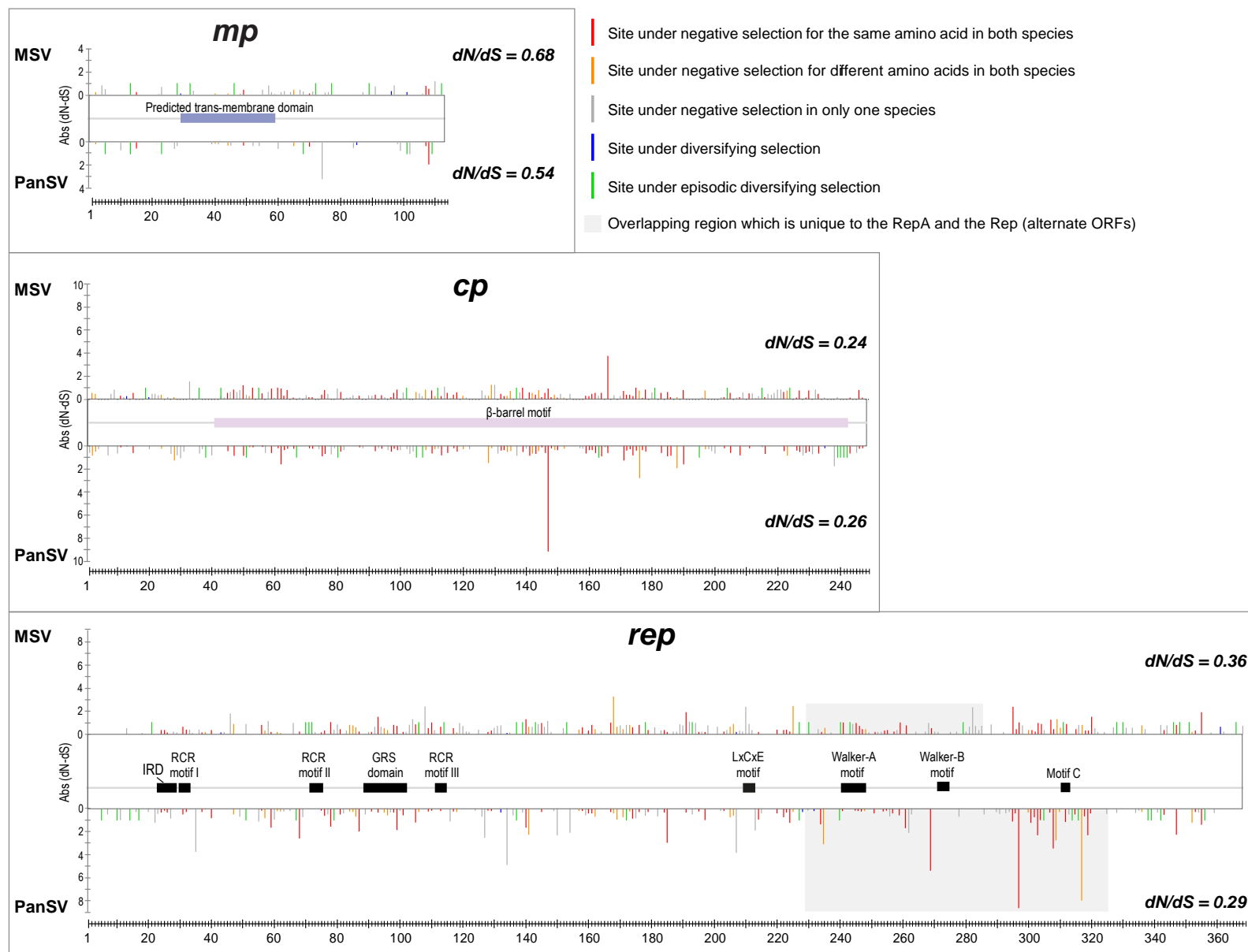


Figure 3.7: See next page for figure legend

Figure 3.7: Plots representing significant signals of natural selection acting on individual codon sites within the *mp*, the *cp* and the *rep* of MSV and PanSV. dN=Non-synonymous substitution rates and dS= Synonymous substitution rates. Absolute (Abs) values of dN-dS are plotted for positive selection (blue) and negative selection (orange, red and grey) indicated for those signals with an associated FUBAR p-value <0.05. Episodic positive selection signals with an associated MEME p-value <0.05 are shown in green. Bar heights for Abs (dN-dS) values correspond to the degree of positive or negative selection detected using FUBAR. Sites at which episodic diversifying selection was detected using MEME have been represented by green bars with uniform height across the genes since Abs (dN-dS) values averaged across the entire phylogeny which do not reflect degrees of episodic diversifying selection (which by definition occurs only on specific subsets of branches within the phylogeny). Total averages for dN/dS ratios are indicated for each gene in MSV and PanSV. Codon sites are indicated based on a codon alignment of both species for each gene. The locations of conserved domains and motifs in relation to their positions in these alignments are shown for the *mp*; the predicted trans-membrane domain (Boulton *et al.*, 1993), the *cp*; the β -barrel motif (Zhang *et al.*, 2001) and for the *rep*; iteron-related domain (IRD) (Argüello-Astorga & Ruiz-Medrano, 2001), rolling circle replication (RCR) motifs I, II and III (Ilyina & Koonin, 1992; Laufs *et al.*, 1995; Rosario *et al.*, 2012), the geminivirus Rep sequence (GRS) domain (Nash *et al.*, 2011), and the helicase domain Walker-A, -B and -C motifs (Gorbalenya & Koonin, 1993; Gorbalenya *et al.*, 1990).

3.5 Concluding remarks

The African monocot-infecting mastreviruses are the most well characterised group of mastreviruses with over 640 genomic sequences recovered from infected grass samples from 18 countries in Africa and four of the surrounding Islands. An overwhelming majority of these are the maize adapted MSV-A which evidently emerged as pathogen of maize ~150 years ago as a result of recombination events among wild grass adapted MSV strains (Harkins *et al.*, 2009; Monjane *et al.*, 2011; Varsani *et al.*, 2008b). Given the extent of recombination detected in MSVs, it is essential to monitor these monocot-infecting mastreviruses infecting wild grasses in order to identify new strains that may pose a significant threat to cultivated grasses. In this study we recovered and sequenced 120 full mastrevirus genomes from predominantly wild uncultivated grasses collected in five countries in Africa and three of the surrounding islands. Our analysis of the African monocot-infecting mastreviruses builds on analyses undertaken in previous studies (Oluwafemi *et al.*, 2011; Oluwafemi *et al.*, 2014; Shepherd *et al.*, 2008b; Varsani *et al.*, 2009; Varsani *et al.*, 2008b) further illuminating the geographic distribution, host range and complex evolutionary dynamics of these viruses.

It is obvious that some species such as MSV and PanSV are widely distributed throughout Africa and neighbouring islands and we now know that MSV has a distribution which extends as far north as Gran Canaria Island. Prior to this study MSV-A and a single isolate of MSV-B had been recovered from maize, however, here we recovered 12 MSV-C genomes from maize plants sampled in Kenya. Several other MSV and PanSV strains were identified to have broader host ranges than previously known with a total of 20 new host genera identified for the various strains. Considering the broad host range of these two mastrevirus species in the context of all other monocot-infecting mastrevirus, all of which have been identified in a three hosts at the most, it is not surprising that they also have moved extensively throughout the continent. The movement and spread of leafhopper vectors and human mediated movement of infected plant material is the most likely route of mastrevirus dispersal throughout the many regions of Africa where these viruses are found.

It is well documented that geminiviruses are able to evolve rapidly through the mechanism of evolution (Kraberger *et al.*, 2012; LaBelle & Gerba, 1980; Lefeuvre *et al.*, 2009; Monjane *et al.*, 2011; Varsani *et al.*, 2009; Varsani *et al.*, 2008a). Our recombination analysis highlights that this mechanism is extremely common in the monocot-infecting mastreviruses. We identified patterns of recombination which are highly conserved in the monocot-infecting mastreviruses and somewhat in geminiviruses. Interestingly we identified the first evidence of ancestral recombination having occurred between monocot-infecting mastrevirus species from two geographically distant regions, Africa and Australia. Further, many of these events involve the exchange of genetic material between species that are not known to infect the same host. Taken together these findings indicate that ancestors of these species at one time most likely occupied the same geographic region(s) and host(s).

For both MSV and PanSV a large amount of sequence data is available and these two species harbour a similar level of diversity which has enabled us to compare signals of selective pressures acting on codon sites within the *mp*, *cp* and *rep* of these species. All genes were evolving under predominantly purifying selection in both species which is common throughout ssDNA viruses (Duffy & Holmes, 2009; Duffy *et al.*, 2008; Hadfield *et al.*, 2012; Kraberger *et al.*, 2012; Stenzel *et al.*, 2014). Notable are the number of sites evolving under apparent episodic diversifying selection, these seem to be scattered throughout the *cp* and *rep* and as sites favouring changes in certain groups of the population these may be a clue to sites that are involved in host specificity.

Overall this study has extended current knowledge of the dynamics of monocot-infecting mastreviruses sampled in Africa and provides some insight into evolutionary forces driving much of these dynamics. In turn some of the mastrevirus species infecting wild uncultivated grasses are apparently wide spread and have been able to adapt to infect a broad range of hosts giving weight to it being equally important to monitor viral populations which infect weed species such as grasses as it is to monitor those that infect crop species.

GenBank accession numbers: KM229914 – KM230033

3.6 References

- Argüello-Astorga, G. R. & Ruiz-Medrano, R. (2001).** An iteron-related domain is associated to Motif 1 in the replication proteins of geminiviruses: Identification of potential interacting amino acid-base pairs by a comparative approach. *Arch Virol* **146**, 1465-1485.
- Bennett, A., McKenna, R. & Agbandje-McKenna, M. (2008).** A comparative analysis of the structural architecture of ssDNA viruses. *Computational and Mathematical Methods in Medicine* **9**, 183-196.
- Bigarré, L., Salah, M., Granier, M., Frutos, R., Thouvenel, J. C. & Peterschmitt, M. (1999).** Nucleotide sequence evidence for three distinct sugarcane streak mastreviruses. *Arch Virol* **144**, 2331-2344.
- Boni, M. F., Posada, D. & Feldman, M. W. (2007).** An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176**, 1035-1047.
- Boulton, M. I., Pallaghy, C. K., Chatani, M., MacFarlane, S. & Davies, J. W. (1993).** Replication of Maize Streak Virus Mutants in Maize Protoplasts: Evidence for a Movement Protein. *Virology* **192**, 85-93.
- Candresse, T., Filloux, D., Muhire, B., Julian, C., Galzi, S., Fort, G., Bernardo, P., Daugrois, J.-H., Fernandez, E., Martin, D. P., Varsani, A. & Roumagnac, P. (2014).** Appearances Can Be Deceptive: Revealing a Hidden Viral Infection with Deep Sequencing in a Plant Quarantine Context. *PLoS ONE* **9**, e102945.
- Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. (2012).** jModelTest 2: more models, new heuristics and parallel computing. *Nature methods* **9**, 772-772.
- Delpont, W., Poon, A. F. Y., Frost, S. D. W. & Kosakovsky Pond, S. L. (2010).** Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**, 2455-2457.
- Duffy, S. & Holmes, E. C. (2009).** Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *Journal of general virology* **90**, 1539-1547.
- Duffy, S., Shackelton, L. A. & Holmes, E. C. (2008).** Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics* **9**, 267-276.
- Edgar, R. C. (2004).** MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797.
- Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. (2000).** Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573-582.
- Giussani, L. M., Cota-Sánchez, J. H., Zuloaga, F. O. & Kellogg, E. A. (2001).** A molecular phylogeny of the grass subfamily Panicoideae (*Poaceae*) shows multiple origins of C4 photosynthesis. *American Journal of Botany* **88**, 1993-2012.
- Gorbalenya, A. E. & Koonin, E. V. (1993).** Helicases: amino acid sequence comparisons and structure-function relationships. *Current Opinion in Structural Biology* **3**, 419-429.

- Gorbalenya, A. E., Koonin, E. V. & Wolf, Y. I. (1990).** A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. *FEBS letters* **262**, 145-148.
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010).** New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**, 307-321.
- Hadfield, J., Thomas, J. E., Schwinghamer, M. W., Kraberger, S., Stainton, D., Dayaram, A., Parry, J. N., Pande, D., Martin, D. P. & Varsani, A. (2012).** Molecular characterisation of dicot-infecting mastreviruses from Australia. *Virus Research* **166**, 13-22.
- Harkins, G. W., Martin, D. P., Duffy, S., Monjane, A. L., Shepherd, D. N., Windram, O. P., Owor, B. E., Donaldson, L., van Antwerpen, T., Sayed, R. A., Flett, B., Ramusi, M., Rybicki, E. P., Peterschmitt, M. & Varsani, A. (2009).** Dating the origins of the maize-adapted strain of maize streak virus, MSV-A. *Journal of General Virology* **90**, 3066-3074.
- Holmes, E. C., Worobey, M. & Rambaut, A. (1999).** Phylogenetic evidence for recombination in dengue virus. *Molecular biology and evolution* **16**, 405-409.
- Hughes, F., Rybicki, E. & Kirby, R. (1993).** Complete nucleotide sequence of sugarcane streak Monogeminivirus. *Arch Virol* **132**, 171-182.
- Ilyina, T. V. & Koonin, E. V. (1992).** Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. *Nucleic Acids Research* **20**, 3279-3285.
- Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. W. (2006).** GARD: a genetic algorithm for recombination detection. *Bioinformatics* **22**, 3096-3098.
- Kraberger, S., Harkins, G. W., Kumari, S. G., Thomas, J. E., Schwinghamer, M. W., Sharman, M., Collings, D. A., Briddon, R. W., Martin, D. P. & Varsani, A. (2013).** Evidence that dicot-infecting mastreviruses are particularly prone to inter-species recombination and have likely been circulating in Australia for longer than in Africa and the Middle East. *Virology* **444**, 282-291.
- Kraberger, S., Thomas, J. E., Geering, A. D. W., Dayaram, A., Stainton, D., Hadfield, J., Walters, M., Parmenter, K. S., van Brunschot, S., Collings, D. A., Martin, D. P. & Varsani, A. (2012).** Australian monocot-infecting mastrevirus diversity rivals that in Africa. *Virus Research* **169**, 127-136.
- LaBelle, R. L. & Gerba, C. P. (1980).** Influence of estuarine sediment on virus survival under field conditions. *Appl Environ Microbiol* **39**, 749-755.
- Laufs, J., Schumacher, S., Geisler, N., Jupin, I. & Gronenborn, B. (1995).** Identification of the nicking tyrosine of geminivirus Rep protein. *FEBS letters* **377**, 258-262.
- Lawry, R., Martin, D., Shepherd, D., van Antwerpen, T. & Varsani, A. (2009).** A novel sugarcane-infecting mastrevirus from South Africa. *Arch Virol* **154**, 1699-1703.
- Lefevre, P., Lett, J.-M., Varsani, A. & Martin, D. (2009).** Widely conserved recombination patterns among single-stranded DNA viruses. *Journal of virology* **83**, 2697-2707.

- Lefevre, P., Martin, D. P., Hoareau, M., Naze, F., Delatte, H., Thierry, M., Varsani, A., Becker, N., Reynaud, B. & Lett, J.-M. (2007).** Begomovirus ‘melting pot’ in the south-west Indian Ocean islands: molecular diversity and evolution through recombination. *Journal of General Virology* **88**, 3458-3468.
- Lima, A. T., Sobrinho, R. R., González-Aguilera, J., Rocha, C. S., Silva, S. J., Xavier, C. A., Silva, F. N., Duffy, S. & Zerbini, F. M. (2013).** Synonymous site variation due to recombination explains higher genetic variability in begomovirus populations infecting non-cultivated hosts. *Journal of General Virology* **94**, 418-431.
- Martin, D. & Rybicki, E. (2000).** RDP: Detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562-563.
- Martin, D. P., Briddon, R. W. & Varsani, A. (2011a).** Recombination patterns in dicot-infecting mastreviruses mirror those found in monocot-infecting mastreviruses. *Arch Virol* **156**, 1463-1469.
- Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D. & Lefevre, P. (2010).** RDP3: A flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**, 2462-2463.
- Martin, D. P., Linderme, D., Lefevre, P., Shepherd, D. N. & Varsani, A. (2011b).** Eragrostis minor streak virus: an Asian streak virus in Africa. *Arch Virol* **156**, 1299-1303.
- Martin, D. P., Posada, D., Crandall, K. A. & Williamson, C. (2005).** A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Research and Human Retroviruses* **21**, 98-102.
- Martin, D. P., Willment, J. A., Billharz, R., Velders, R., Odhiambo, B., Njuguna, J., James, D. & Rybicki, E. P. (2001).** Sequence diversity and virulence in Zea mays of Maize streak virus isolates. *Virology* **288**, 247-255.
- Monjane, A. L., Harkins, G. W., Martin, D. P., Lemey, P., Lefevre, P., Shepherd, D. N., Oluwafemi, S., Simuyandi, M., Zinga, I., Komba, E. K., Lakoutene, D. P., Mandakombo, N., Mboukoulida, J., Semballa, S., Tagne, A., Tiendrebeogo, F., Erdmann, J. B., van Antwerpen, T., Owor, B. E., Flett, B., Ramusi, M., Windram, O. P., Syed, R., Lett, J. M., Briddon, R. W., Markham, P. G., Rybicki, E. P. & Varsani, A. (2011).** Reconstructing the history of maize streak virus strain a dispersal to reveal diversification hot spots and its origin in southern Africa. *Journal of Virology* **85**, 9623-9636.
- Muhire, B., Martin, D. P., Brown, J. K., Navas-Castillo, J., Moriones, E., Zerbini, M. F., Rivera-Bustamante, R. F., Malathi, V. G., Briddon, R. W. & Varsani, A. (2013).** A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). *Arch Virol* **158**, 1411-1424.
- Muhire, B. M., Varsani, A. & Martin, D. P. (2014).** SDT: A Virus Classification Tool Based on Pairwise Sequence Alignment and Identity Calculation. *PLoS ONE* **9**, e108277.
- Mullineaux, P., Donson, J., Morris-Krsinich, B., Boulton, M. & Davies, J. (1984).** The nucleotide sequence of maize streak virus DNA. *The EMBO journal* **3**, 3063.
- Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Pond, S. L. K. & Scheffler, K. (2013).** FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Molecular biology and evolution* **30**, 1196-1205.

- Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K. & Pond, S. L. K. (2012).** Detecting individual sites subject to episodic diversifying selection. *Plos Genet* **8**, e1002764.
- Nash, T. E., Dallas, M. B., Reyes, M. I., Buhrman, G. K., Ascencio-Ibañez, J. T. & Hanley-Bowdoin, L. (2011).** Functional analysis of a novel motif conserved across geminivirus Rep proteins. *Journal of Virology* **85**, 1182-1192.
- Olmstead, R. G. & Reeves, P. A. (1995).** Evidence for the polyphyly of the Scrophulariaceae based on chloroplast *rbcL* and *ndhF* sequences. *Annals of the Missouri Botanical Garden* **82**, 176-193.
- Oluwafemi, S., Alegbejo, M. D., Onasanya, A. & Olufemi, O. (2011).** Relatedness of Maize streak virus in maize (*Zea mays* L.) to some grass isolates collected from different regions in Nigeria. *African Journal of Agricultural Research* **6**, 5878-5883.
- Oluwafemi, S., Kraberger, S., Shepherd, D. N., Martin, D. P. & Varsani, A. (2014).** A high degree of African streak virus diversity within Nigerian maize fields includes a new mastrevirus from *Axonopus compressus*. *Arch Virol* **159**, 2765-2770.
- Oluwafemi, S., Varsani, A., Monjane, A. L., Shepherd, D. N., Owor, B. E., Rybicki, E. P. & Martin, D. P. (2008).** A new African streak virus species from Nigeria. *Arch Virol* **153**, 1407-1410.
- Owor, B. E., Martin, D. P., Shepherd, D. N., Edema, R., Monjane, A. L., Rybicki, E. P., Thomson, J. A. & Varsani, A. (2007).** Genetic analysis of maize streak virus isolates from Uganda reveals widespread distribution of a recombinant variant. *Journal of General Virology* **88**, 3154-3165.
- Padidam, M., Sawyer, S. & Fauquet, C. M. (1999).** Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218-225.
- Pande, D., Kraberger, S., Lefeuvre, P., Lett, J.-M., Shepherd, D., Varsani, A. & Martin, D. (2012).** A novel maize-infecting mastrevirus from La Réunion Island. *Arch Virol* **157**, 1617-1621.
- Peterschmitt, M., Granier, M., Frutos, R. & Reynaud, B. (1996).** Infectivity and complete nucleotide sequence of the genome of a genetically distinct strain of maize streak virus from Reunion Island. *Arch Virol* **141**, 1637-1650.
- Posada, D. & Crandall, K. A. (2001).** Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 13757-13762.
- Rosario, K., Duffy, S. & Breitbart, M. (2012).** A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol* **157**, 1851-1871.
- Sanz, A. I., Fraile, A., García-Arenal, F., Zhou, X., Robinson, D. J., Khalid, S., Butt, T. & Harrison, B. D. (2000).** Multiple infection, recombination and genome relationships among begomovirus isolates found in cotton and other plants in Pakistan. *Journal of General Virology* **81**, 1839-1849.
- Shackelton, L. A., Parrish, C. R., Truyen, U. & Holmes, E. C. (2005).** High rate of viral evolution associated with the emergence of carnivore parvovirus. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 379-384.

- Shepherd, D. N., Martin, D. P., Lefeuvre, P., Monjane, A. L., Owor, B. E., Rybicki, E. P. & Varsani, A. (2008a).** A protocol for the rapid isolation of full geminivirus genomes from dried plant tissue. *Journal of Virological Methods* **149**, 97-102.
- Shepherd, D. N., Martin, D. P., Van Der Walt, E., Dent, K., Varsani, A. & Rybicki, E. P. (2010).** Maize streak virus: An old and complex 'emerging' pathogen. *Molecular Plant Pathology* **11**, 1-12.
- Shepherd, D. N., Varsani, A., Windram, O. P., Lefeuvre, P., Monjane, A. L., Owor, B. E. & Martin, D. P. (2008b).** Novel sugarcane streak and sugarcane streak Reunion mastreviruses from southern Africa and la Réunion. *Arch Virol* **153**, 605-609.
- Silva, F. N., Lima, A. T., Rocha, C. S., Castillo-Urquiza, G. P., Alves-Júnior, M. & Zerbini, F. M. (2014).** Recombination and pseudorecombination driving the evolution of the begomoviruses Tomato severe rugose virus (ToSRV) and Tomato rugose mosaic virus (ToRMV): two recombinant DNA-A components sharing the same DNA-B. *Virology journal* **11**, 66.
- Smith, J. M. (1992).** Analyzing the mosaic structure of genes. *Journal of Molecular Evolution* **34**, 126-129.
- Stenzel, T., Piasecki, T., Chrzastek, K., Julian, L., Muhire, B. M., Golden, M., Martin, D. P. & Varsani, A. (2014).** Pigeon circoviruses display patterns of recombination, genomic secondary structure and selection similar to those of beak and feather disease viruses. *Journal of General Virology* **95**, 1338-1351.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. (2011).** MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**, 2713-2739.
- Varsani, A., Martin, D. P., Navas-Castillo, J., Moriones, E., Hernández-Zepeda, C., Idris, A., Zerbini, F. M. & Brown, J. K. (2014).** Revisiting the classification of curtoviruses based on genome-wide pairwise identity. *Arch Virol* **159**, 1873-1882.
- Varsani, A., Monjane, A. L., Donaldson, L., Oluwafemi, S., Zinga, I., Komba, E. K., Plakoutene, D., Mandakombo, N., Mboukoulida, J., Semballa, S., Briddon, R. W., Markham, P. G., Lett, J. M., Lefeuvre, P., Rybicki, E. P. & Martin, D. P. (2009).** Comparative analysis of Panicum streak virus and Maize streak virus diversity, recombination patterns and phylogeography. *Virology Journal* **6**, 194.
- Varsani, A., Oluwafemi, S., Windram, O., Shepherd, D., Monjane, A., Owor, B., Rybicki, E., Lefeuvre, P. & Martin, D. (2008a).** Panicum streak virus diversity is similar to that observed for maize streak virus. *Arch Virol* **153**, 601-604.
- Varsani, A., Shepherd, D. N., Monjane, A. L., Owor, B. E., Erdmann, J. B., Rybicki, E. P., Peterschmitt, M., Briddon, R. W., Markham, P. G., Oluwafemi, S., Windram, O. P., Lefeuvre, P., Lett, J. M. & Martin, D. P. (2008b).** Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. *Journal of General Virology* **89**, 2063-2074.
- Zhang, W., Olson, N. H., Baker, T. S., Faulkner, L., Agbandje-McKenna, M., Boulton, M. I., Davies, J. W. & McKenna, R. (2001).** Structure of the Maize Streak Virus Geminale Particle. *Virology* **279**, 471-477.

Chapter 4

Evidence that dicot-infecting mastreviruses are particularly prone to inter-species recombination and have likely been circulating in Australia for longer than in Africa and the Middle East

Contents

4.1	Abstract.....	152
4.2	Introduction.....	153
4.3	Materials and methods.....	154
4.3.1	Sample collection, virus isolation and genome cloning	154
4.3.2	Sequence assembly and pairwise sequence analyses	155
4.3.3	Recombination analysis and construction of mostly recombination-free datasets	156
4.3.4	Phylogenetic analyses and identification of the likely origin of dicot-infecting mastreviruses.....	156
4.4	Results and discussion	158
4.4.1	Classification of new dicot-infecting mastrevirus full genome sequences	158
4.4.2	Complex patterns of inter- and intra-species recombination amongst dicot-infecting mastreviruses	163
4.4.3	The geographical origin of the dicot-infecting mastreviruses	168
4.4.4	Plausible routes of dicot infecting mastrevirus movement out of Australia	169
4.5	Concluding remarks.....	176
4.6	References.....	180

This body of work has been published in Virology and is presented in a similar manner to that of the publication:

Kraberger, S., Harkins, G.W., Kumari, S.G., Thomas, J.E., Schwinghamer, M.W., Sharman, M., Collings, D.A., Briddon, R.W., Martin, D.P. and Varsani, A. (2013) Evidence that dicot-infecting mastreviruses are particularly prone to inter-species recombination and have likely been circulating in Australia for longer than in Africa and the Middle East.

Virology 444 (1–2), 282–291.

4.1 Abstract

Viruses of the genus *Mastrevirus* (family *Geminiviridae*) are transmitted by leafhoppers and infect either mono- or dicotyledonous plants. Here we have determined the full length sequences of 49 dicot-infecting mastrevirus isolates sampled in Australia, Eritrea, India, Iran, Pakistan, Syria, Turkey and Yemen. Comprehensive analysis of all available dicot-infecting mastrevirus sequences showed the diversity of these viruses in Australia to be greater than in the rest of their known range, consistent with earlier studies, and that, in contrast with the situation in monocot-infecting mastreviruses, detected inter-species recombination events outnumbered intra-species recombination events. Consistent with Australia having the greatest diversity of known dicot-infecting mastreviruses, phylogeographic analyses indicating the most plausible scheme for the spread of these viruses to their present locations, suggest that most recent common ancestor of these viruses is likely nearer Australia than it is to the other regions investigated.

4.2 Introduction

Chapter Two and Three involved an in depth examination into the monocot-infecting mastreviruses present in two major diversity hotspots, Australia and Africa. The work undertaken in this Chapter examines the dicot-infecting mastreviruses from a global perspective to see if similar patterns are evident and gain insights into possible origins of these viruses.

Throughout the agricultural regions of Australia, south and north-east Africa, the Middle East and India, mastreviruses are recognised as potentially important threats to chickpea (*Cicer arietinum*), lentil (*Lens culinaris*), bean (*Phaseolus vulgaris*) and tobacco (*Nicotiana tabacum*) production (Farzadfar *et al.*, 2002; Hadfield *et al.*, 2012; Halley-Stott *et al.*, 2007; Horn *et al.*, 1994; Horn *et al.*, 1993; Kumari *et al.*, 2004; Kumari *et al.*, 2008; Makkouk *et al.*, 2003; Mumtaz *et al.*, 2011; Nahid *et al.*, 2008; Schwinghamer *et al.*, 2010; Thomas *et al.*, 2010). Besides being economically important export crops for countries such as Australia, pulses such as lentils, chickpeas and beans are key dietary staples in northern Africa, India, Pakistan and the Middle East; with India alone producing around five million tonnes per annum over four decades to 2005 (Knights *et al.*, 2007). By influencing the yields of important food crops in these populous and often agriculturally marginal regions, pathogens including mastreviruses threaten the food security of a substantial number of the world's most economically vulnerable people.

According to the most recent report by the International Committee of the Taxonomy of Viruses: Geminiviridae Study Group on mastrevirus classification there are six known species of dicot-infecting mastreviruses (Muhire *et al.*, 2013). One species, *Chickpea chlorotic dwarf virus* (CpCDV), has been found only in the Middle East (including Turkey), Africa and India (Ali *et al.*, 2004; Horn *et al.*, 1993; Mumtaz *et al.*, 2011; Nahid *et al.*, 2008). All five of the other recognised species have only ever been found in Australia. These include *Chickpea redleaf virus* (CpRLV) (Thomas *et al.*, 2010), *Chickpea yellows virus* (CpYV) (Hadfield *et al.*, 2012), *Chickpea chlorosis virus* (CpCV) (Hadfield *et al.*, 2012; Thomas *et al.*, 2010), *Chickpea chlorosis Australia virus* (CpCAV) (Hadfield *et al.*, 2012) and *Tobacco yellow dwarf virus* (TYDV) (Hadfield *et al.*, 2012; Morris *et al.*, 1992). This known distribution of dicot-infecting mastrevirus species is less extensive than that of the monocot-infecting

mastreviruses, which have been identified in Africa, Europe, Asia, Indian Ocean islands, throughout the Pacific rim and, more recently, in the Caribbean (Muhire *et al.*, 2013; Rosario *et al.*, 2013).

Consistent with the notion that Australia is both the present centre of dicot-infecting mastrevirus diversity and is close to the region where these viruses first emerged, is that CpCDV is the only dicot-infecting mastrevirus species to be discovered outside of Australia, and phylogenetic evidence indicating that CpCDV forms a distinct monophyletic clade with high statistical support that it is nested within a much larger clade that contains the five Australian species (Hadfield *et al.*, 2012). It is, however, also possible that this view of dicot-infecting mastrevirus diversity has been biased by the fact that Australia is the site where these viruses have been most intensively sampled. Also it is entirely plausible that, as more dicot-infecting mastreviruses are sampled from elsewhere in the world, a completely different picture will emerge.

In order to get a better perspective of the extent of dicot-infecting mastrevirus diversity in other parts of the world, we determined the full genome sequences of 30 isolates from symptomatic leaf material collected in north-east Africa, the Middle East (including Turkey) and India between 1993 and 2005. We also determined the full genome sequences of 19 dicot-infecting mastrevirus isolates recovered from symptomatic plant samples collected in Australia between 2002 and 2011. This dataset was analysed together with all previously described monocot- and dicot-infecting mastreviruses and through this we identified six divergent strains of CpCDV.

4.3 Materials and methods

4.3.1 Sample collection, virus isolation and genome cloning

Samples from 49 pulses chickpea (*Cicer arietinum*), lentil (*Lens culinaris*), faba bean (*Vicia faba*), field pea (*Pisum sativum*) and bean (*Phaseolus vulgaris*), collected in Syria (n=2), Pakistan (n=1), India (n=2), Turkey (n=2), Eritrea (n=9), Iran (n=9), Yemen (n=5) and Australia (n=19) which had previously been identified to be positive for mastreviruses either by PCR or ELISA were used in this study (Additional Table 4.1 details host species for each

sample). Total DNA was extracted from plant sap or dried plant material using Epoch nucleic acid purification kits (Epoch Life Science, USA). Enrichment of circular viral DNA from total DNA was carried out using the Illustra TempliPhi Amplification Kit (GE Healthcare, USA) as previously described by Owor *et al.* (2007) and Shepherd *et al.* (2008). Viral DNA amplicons were then digested using the restriction enzymes *Hind*III or *Xmn*I which yielded ~2.6 kb linearised unit length genomes. These were gel purified and ligated at either the *Hind*III or *Xmn*I sites of the cloning vector, pGEM3Zf+ (Promega Biotech, USA).

We used a polymerase chain reaction (PCR) amplification approach to recover viral genomes from 44 of the 49 TempliPhi enriched DNA samples for which we were unable to find a unique restriction enzyme. Degenerate back-to-back primers (dicot forward 5'-GAN TTG GTC CGC AGT GTA GA-3', dicot reverse 5'-GTA CCG GWA AGA CMW CYT GG-3'), previously described by Hadfield *et al.* (2012) were used to amplify full length dicot-infecting mastrevirus genomes using Kapa HiFi HotStart DNA polymerase (Kapa Biosystems, USA) with the following thermocycling conditions: 94°C for 3 min, 25 cycles of 98°C (3 min), 52°C (30 sec), 72°C (2.45 min) and a final extension of 72°C for 3 min. PCR amplicons were ligated into linearised pJET1.2 vector (CloneJET™ PCR cloning kit, Fermentas, USA). All plasmids with cloned viral genomes were sequenced at MacroGen (Korea) by primer walking.

4.3.2 Sequence assembly and pairwise sequence analyses

Viral genome sequences were assembled using DNAMAN (version 7; Lynnon Biosoft, Canada). Forty-eight dicot-infecting mastrevirus full genome sequences available in public databases on 24 October 2012 and the wheat dwarf virus sequence (AM040732; included as an outlier) were obtained and aligned with the sequences determined in this study using MUSCLE (Edgar, 2004). The nucleotide sequence alignment thus obtained was manually edited using MEGA5 (Tamura *et al.*, 2011). Similarly, putative Rep, MP and CP encoding sequences of the 97 virus genomes were computationally translated and aligned using MEGA5 with manual editing. Pairwise identities (1 - p-distance, with pairwise deletion of gaps) of the full dicot-infecting mastrevirus genomes were determined using SDT v1.0 (Muhire *et al.*, 2013).

4.3.3 Recombination analysis and construction of mostly recombination-free datasets

Recombination analysis within the dicot-infecting mastreviruses was performed using RDP4 (Martin *et al.*, 2010), with the following methods; RDP, GENECONV (Padidam *et al.*, 1999), Bootscan (Martin *et al.*, 2005), Maxchi (Smith, 1992), Chimera (Posada & Crandall, 1998), Siscan (Gibbs *et al.*, 2000), and 3Seq (Boni *et al.*, 2007). Potential recombination signals were accepted as being genuine evidence of actual recombination events when they were detected with three or more of the seven methods (with associated p-values of $<10^{-3}$) coupled with phylogenetic support for recombination having occurred.

Based on the recombination analysis, two mostly recombination-free sequence alignments - corresponding to a coat protein (CP) gene dataset and a Rep gene dataset were extracted from the full genome sequence alignments.

4.3.4 Phylogenetic analyses and identification of the likely origin of dicot-infecting mastreviruses

A maximum likelihood (ML) phylogenetic tree of the aligned full genome sequences (with recombinant region removed) was constructed using PHYML version 3 (Guindon *et al.*, 2010) with 1000 non-parametric bootstrap replicates with GTR+G4 selected as the best fit nucleotide substitution model using RDP 4 (Martin *et al.*, 2010) and rooted with *Wheat dwarf virus* (WDV). Branches with less than 60% bootstrap support were manually collapsed using MESQUITE (Version 2.75).

We opted to use Bayesian maximum clade credibility (MCC) trees produced using the computer program BEAST (Drummond *et al.*, 2012) to evaluate the likely geographical origin of the dicot-infecting mastreviruses. These trees were time-calibrated based on sequence sampling times, with the root location based on the most plausible dating of the most recent common ancestor (MRCA) of the analysed sequences. Each of the MCC trees produced by BEAST represented an entire distribution of similarly plausible trees and explicitly accounted for phylogenetic uncertainty during their inference. Furthermore, besides offering fully probabilistic models of sequence evolution, BEAST also implements phylogeographic models of sequence movement between discrete sampling locations (such as between cities, provinces, countries or other discrete geographical regions). These models have been employed previously to investigate the movement dynamics the monocot-infecting mastrevirus species, *Maizke streak virus* (Monjane *et al.*, 2011) and the begomovirus

species': *Tomato yellow leaf curl virus* (Lefeuvre *et al.*, 2010) and *East African cassava mosaic virus* (De Bruyn *et al.*, 2012). The discrete phylogeography model used here to infer when and where the MRCA of the dicot-infecting mastreviruses existed considered geographic diffusion among six discrete sample locations: the Western Mediterranean (WM), Asia (AS) the Middle East (ME), East Africa (EA), Southern Africa (SA) and Australia (AU).

Since previous analyses have indicated that sampling biases can strongly influence the phylogeographic inference of ancestral sequence locations in BEAST (De Bruyn *et al.*, 2012; Lefeuvre *et al.*, 2010; Monjane *et al.*, 2011) we took steps to both directly reduce the influences of these biases prior to analyses, and to test for the effects of any biases after the analyses were concluded. Specifically, we randomly removed all but 10 of the Australian sequences from the full genome, CP and Rep datasets pre-analysis. Post-analysis, we directly evaluated the effects of residual sampling biases on the inferred geographical location of the MRCA by randomly swapping sampling locations among the sequences followed by revaluation of the MRCA location state. This test would indicate that a sampling bias had influenced inference of the MRCA location if the same location(s) were indicated for the MRCA in both the randomised and un-randomised analyses.

For each of the analysed datasets, independent replicate runs of the Markov chain of 2×10^7 steps were performed using BEAST so as to achieve effective sample size (ESS) estimates for all relevant model parameters that were always >200 .

The degree of clock-like evolution evident within the analysed sequence datasets (full genome, CP and Rep) was evaluated using root-to-tip genetic distance vs. sampling date regression analyses based on inferred neighbor-joining trees using the computer program, Path-O-Gen (available from <http://tree.bio.ed.ac.uk/software/pathogen/>) (Drummond *et al.*, 2003).

We used the computer program SPREADv1.0.4 (Bielejec *et al.*, 2011) (available from www.kuleuven.ac.be/aidslab/phylogeography/SPREAD.html) to perform Bayes factor (BF) tests of potential epidemiological links between the analysed geographical regions revealed by the phylogeographic analyses performed by BEAST. In these tests we accepted $BF_{\log10}$ values greater than or equal to 5.0 as being indicative of significant statistical support for

movement between pairs of geographical regions (where a $BF_{\log10} > 100$ was taken to represent decisive support, a $BF_{\log10} > 10.0$ was taken to represent strong support and a $BF_{\log10} < 5.0$ was taken to represent poor support.). SPREAD was then used to produce .kml formatted files containing information on BF test supported routes of virus movement. These files can be viewed using the computer program, Google Earth (available from <http://earth.google.com>).

4.4 Results and discussion

4.4.1 Classification of new dicot-infecting mastrevirus full genome sequences

Forty-nine dicot-infecting mastrevirus genomes (Table 1) were recovered from chickpea (n=40), lentil (n=4), faba bean (n=2), field pea (n=2) and bean (n=1). These 49 viral genomes and 48 others available in GenBank were assembled into a single dataset and genome-wide pairwise identities between every possible pair of sequences were calculated (1 minus *p*-distance calculated with pairwise deletion of gaps; Fig. 4.1A) so as to assess the over-all genetic diversity of these viruses. Based on the recommendations of Muhire et al (2013) eighteen of the nineteen Australian dicot-infecting mastrevirus genomes could be assigned to previously named species and strain groupings; TYDV (1/19), CpCAV (7/19), CpCV-A (3/19), CpCV-B (1/19), and CpCV-E (6/19). The one exceptional Australian dicot-infecting isolate was clearly a member of the species CpCV but was <87% similar to any previously described CpCV isolate and was therefore assigned to a new strain of this species: CpCV-F. The 30 dicot-infecting mastreviruses from north-east Africa, the Middle East and the Indian subcontinent are all CpCDV isolates, either classifiable as members of the previously described CpCDV strains -A (11/30), and -D (2/30), or, because they shared < 94% identity to isolates in previously described strains, were assigned to new strains -F (8/30), -G (2/30), -H (1/30), -I (1/30), -J (1/30) and -K (4/30).

It is evident both from the identity scores of all pairs of available dicot-infecting mastrevirus sequences and the maximum identity scores of all pairs of isolates within individual species, that even within individual species there is greater diversity amongst the known Australian dicot-infecting mastrevirus isolates than there is amongst the CpCDV isolates found across north-east Africa, South Africa, the Middle East, Turkey, Pakistan and India combined (Fig. 4.1).

Table 4.1: Details for all full dicot-infecting mastrevirus genomes available in GenBank, including those from this study. GenBank accessions in bold are those genomes determined in this study.

Species	Strain	GenBank no.	Country	Host	Sampling year
CpCDV	CpCDV-A	FR687959	Syria	Chickpea (<i>Cicer arietinum</i>)	2008
		KC172662	Turkey	Chickpea	1996
		KC172663	Turkey	Chickpea	1996
		KC172655	Iran	Chickpea	1999
		KC172653	Iran	Chickpea	1999
		KC172654	Iran	Chickpea	2002
		KC172656	Iran	Chickpea	1999
		KC172657	Iran	Chickpea	1999
		KC172658	Iran	Chickpea	1999
		KC172659	Iran	Chickpea	1999
		KC172660	Iran	Chickpea	1999
		KC172661	Iran	Field Pea (<i>Pisum sativum</i>)	1999
	CpCDV-B	Y11023	South Africa	Bean (<i>Phaseolus vulgaris</i>)	1997
		DQ458791	South Africa	Bean	1997
		AM849096	Pakistan	Chickpea	2005
	CpCDV-C	AM849097	Pakistan	Chickpea	2005
		AM850136	Pakistan	Chickpea	2007
		AM900416	Pakistan	Chickpea	2007
	CpCDV-D	FR687960	Pakistan	Chickpea	2008
		KC172664	India	Chickpea	1993
		KC172665	India	Field Pea	1993
	CpCDV-E	AM933135	Sudan	Chickpea	1997
		AM933134	Sudan	Chickpea	1997
	CpCDV-F	KC172666	Pakistan	Lentil (<i>Lens culinaris</i>)	1997
		KC172669	Yemen	Lentil	1996
		KC172672	Yemen	Lentil	1996
		KC172673	Yemen	Lentil	1996
		KC172670	Yemen	Faba bean (<i>Vicia faba</i>)	1996
		KC172671	Yemen	Faba bean	1996
		KC172667	Syria	Chickpea	2003
		KC172668	Syria	Chickpea	1999
	CpCDV-G	KC172674	Eritrea	Chickpea	2005
		KC172675	Eritrea	Chickpea	2005
	CpCDV-H	KC172676	Eritrea	Chickpea	2005
	CpCDV-I	KC172677	Eritrea	Chickpea	2005
	CpCDV-J	KC172678	Eritrea	Chickpea	2005
	CpCDV-K	KC172679	Eritrea	Chickpea	2005
		KC172680	Eritrea	Chickpea	2005
		KC172681	Eritrea	Chickpea	2005
		KC172682	Eritrea	Chickpea	2005
CpCV	CpCV-A	GU256530	Australia	Chickpea	2002
		JN989413	Australia	Chickpea	2002
		JN989414	Australia	Chickpea	2002
		JN989415	Australia	Chickpea	2002
		KC172685	Australia	Chickpea	2010
		KC172683	Australia	Chickpea	2002
		KC172684	Australia	Chickpea	2002
	CpCV-B	GU256531	Australia	Chickpea	2003
		KC172690	Australia	Chickpea	2011
	CpCV-C	JN989416	Australia	Chickpea	2002
		JN989417	Australia	Chickpea	2002

Table 4.1 continued

Species	Strain	GenBank no.	Country	Host	Sampling year
CpCV-E		JN989438	Australia	Bean	1984
		JN989426	Australia	Chickpea	2002
		JN989437	Australia	Chickpea	2002
		JN989429	Australia	Chickpea	2002
		JN989434	Australia	Chickpea	2002
		JN989428	Australia	Chickpea	2002
		JN989430	Australia	Chickpea	2002
		JN989431	Australia	Chickpea	2002
		JN989432	Australia	Chickpea	2002
		JN989433	Australia	Chickpea	2002
		KC172699	Australia	Chickpea	2002
		KC172698	Australia	Chickpea	2002
		KC172694	Australia	Chickpea	2002
		KC172695	Australia	Chickpea	2002
		KC172696	Australia	Chickpea	2002
		KC172697	Australia	Chickpea	2002
CpCV-F		KC172700	Australia	Chickpea	2002
CpCAV		JN989418	Australia	Bean	2007
		JN989419	Australia	Chickpea	2010
		JN989420	Australia	Chickpea	2010
		JN989421	Australia	Chickpea	2010
		JN989422	Australia	Chickpea	2002
		JN989423	Australia	Chickpea	2003
		KC172691	Australia	Chickpea	2011
		KC172693	Australia	Chickpea	2011
		KC172692	Australia	Chickpea	2011
		KC172689	Australia	Chickpea	2002
		KC172686	Australia	Chickpea	2003
		KC172687	Australia	Chickpea	2003
		KC172688	Australia	Chickpea	2003
CpRLV		GU256532	Australia	Chickpea	2003
CpYV		JN989439	Australia	Chickpea	2002
TYDV		M81103	Australia	Tobacco (<i>Nicotinana sp.</i>)	1992
		JN989440	Australia	Tobacco	1986
		JN989445	Australia	Tobacco	1985
		JN989446	Australia	Tobacco	2002
		JN989441	Australia	Bean	2010
		JN989442	Australia	Bean	2010
		JN989443	Australia	Bean	2010
		KC172702	Australia	Bean	2010
		JN989444	Australia	Chickpea	2002

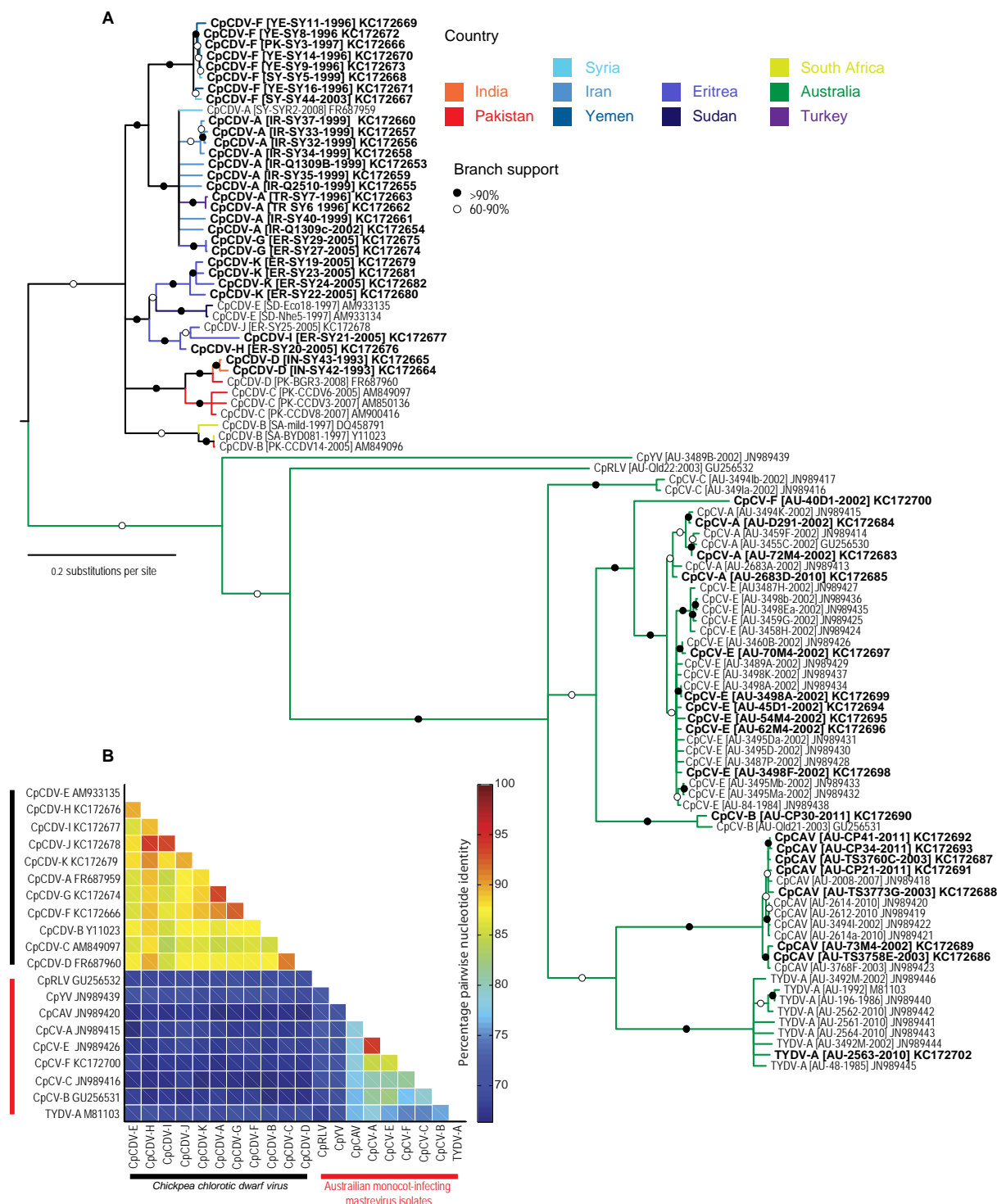


Figure 4.1: (A) Maximum likelihood phylogenetic tree (constructed with the nucleotide substitution model GTR+G4) of all available dicot-infecting mastrevirus full genome sequences (with recombinant regions removed). The trees were rooted with WDV. Bootstrap support for branches is indicated by open (60-89%) and closed circles (>90%), branches with less than 60% bootstrap support have been collapsed. Countries of origin are represented by colours shown in key. Viral isolate sequences determined in this study have accession numbers KC172653 – KC172702. (B) Two dimensional percentage pairwise identity plot matrix of a representative dicot-infecting mastrevirus from each strain and species.

4.4.2 Complex patterns of inter- and intra-species recombination amongst dicot-infecting mastreviruses

As has been demonstrated previously with smaller datasets, recombination has played a major role in the evolution of dicot-infecting mastreviruses (Hadfield *et al.*, 2012; Martin *et al.*, 2011b). A total of 16 intra-species and 10 inter-species recombination events were detected. Although 12 of the recombination events detected here were previously identified by Martin *et al.* (2011b) and Hadfield *et al.* (2012), the additional full genome sequences generated during this study has increased the resolution with which many of these recombination events can be characterised (Fig. 4.2; Table 4.2).

Several groups of isolates apparently carry evidence of multiple independent recombination events. For example, the CpCV-F isolate has evidence of one intra-species recombination event involving the acquisition by an ancestral CpCV-E-like virus of a *cp* gene fragment from a CpCV-C-like virus (Event 1 in Fig. 4.2; Table 4.2). The ancestral CpCV-E-like sequence from which the ancestor of the CpCV-F sequences was likely derived was, as is the case with all contemporary CpCV-E and CpCV-A sequences, in turn carrying evidence of a likely much older inter-species recombination event. These events involved the transfer of a *rep* gene fragment from a CpCAV-like sequence into the genome of a CpCV-B-like sequence (Event F in Fig. 4.2; Table 4.2). More recently than the two previously discussed events detectable within the CpCV-F sequences, was an event involving a small region of the SIR of a common ancestor of these sequences which appears to have been derived from a currently unknown monocot-infecting mastrevirus species (Event G in Fig. 4.2; Table 4.2). Similarly complex recombination patterns are detectable within the sampled CpCV-A and CpCDV-K genomes, suggesting that such convoluted evolutionary histories might be fairly common amongst the dicot-infecting mastreviruses.

Consistent with previous analyses of the monocot-infecting mastreviruses, we detected (1) that intra-species recombination events, in most cases, have tended to involve transfers of larger genome fragments (average of 22% ranging between 10% and 49% of the genome) than inter-species recombination events (average of 17% ranging between 10% and 30% of the genome; (Martin *et al.*, 2001; Varsani *et al.*, 2009a; Varsani *et al.*, 2008b) and (2) that there are clear recombination breakpoint hotspots within the LIR and SIR genome regions (Martin *et al.*, 2011b), and (3) a greater number of recombination breakpoints in the

complementary sense genes than in the virion sense genes (Hadfield *et al.*, 2012; Krabberger *et al.*, 2012; Martin *et al.*, 2011b; Owor *et al.*, 2007; Varsani *et al.*, 2009a; Varsani *et al.*, 2008b). The concentration of recombination breakpoints within the intergenic regions of these viruses enabled us to construct two relatively recombination-free datasets corresponding to the *cp* and *rep* gene regions of the full genome dataset – hereafter respectively referred to as the CP and Rep datasets.

It is interesting to note that recombination events were detected between species from two geographically separated regions, those species found in Australia and CpCDV which has only been documented in Africa, the Middle East and Indian Subcontinent. Evidence of similar recombination events in the monocot-infecting mastreviruses discussed in Chapter Two is a strong indication that these mastrevirus species likely circulated in the same region and infected the same host(s) at some point.

Inter- and intra-species recombination

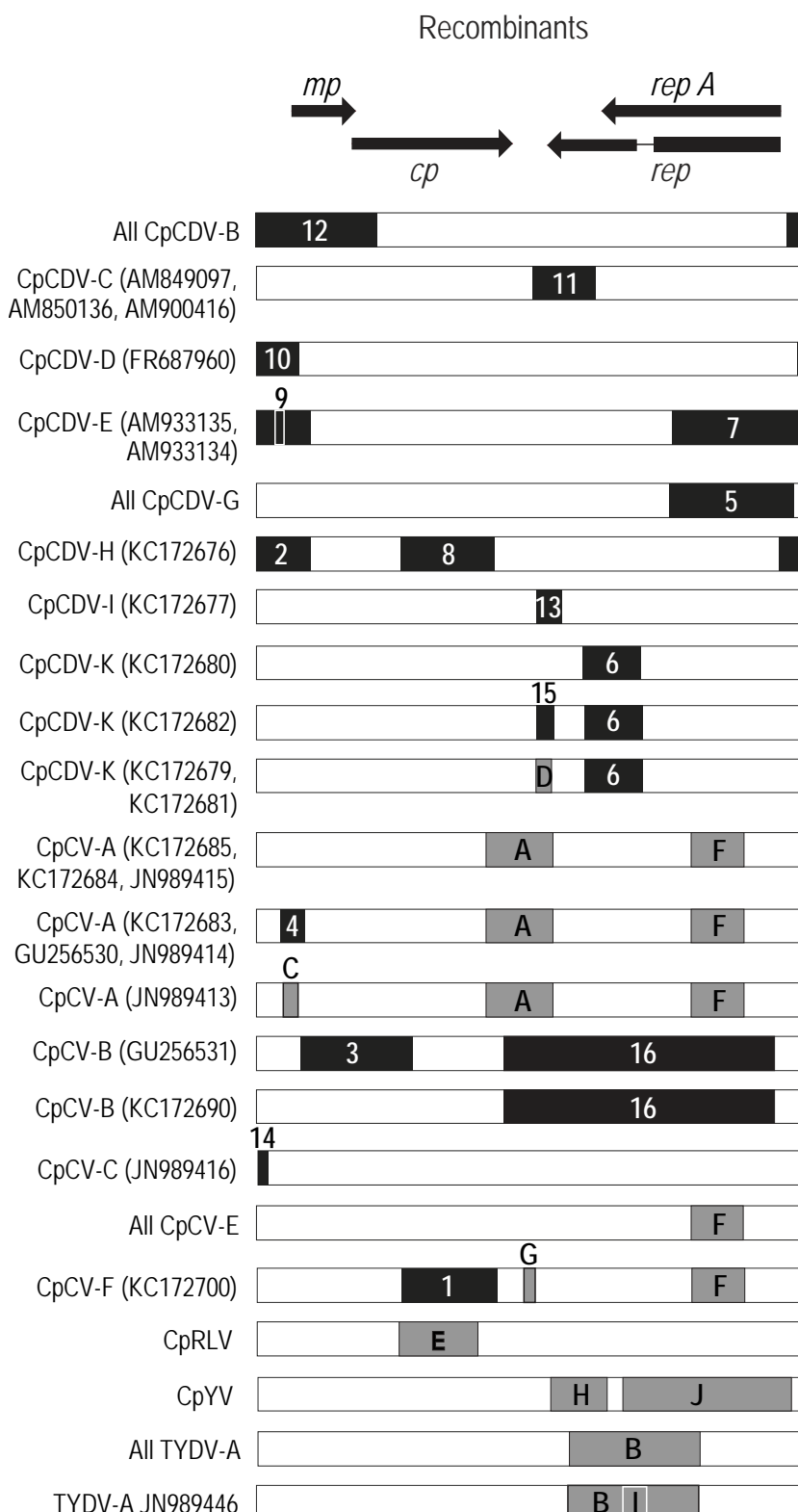


Figure 4.2: Illustration of recombination events amongst all dicot-infecting mastrevirus isolates. Inter-species recombination events are represented in grey and have an associated letter code. Intra-species events are represented in black and have an associated number code. Arrows above the genome maps indicate the positions of the *mp* (movement protein), *cp* (coat protein), *repA* (replication-associated A protein) and *rep* (replication-associated protein) genes.

Table 4.2: Details of all recombination events detected. Major and minor parents are inferred based on genetic fragments they donated to the recombinant, with the major parent donating the larger fragment and the minor parent the smaller fragment. The method with the most significant associated p-value is indicated in bold for each event.

Event	Recombinant region	Potential major Parent	Potential minor Parent	Detection method	P-value
Intra-species recombination					
1	646-1127	All CpCV-E	CpCV-C (JN989416, JN989417)	RGMCT	7.45×10^{-21}
2	2465-212	CpCDV-J (KC172678)	Unknown	RGMCT	2.24×10^{-12}
3	141-697	CpCV-B (KC172690)	CpCV-A (KC172683, KC172684, KC172685, GU256530, JN989413, JN989414, JN989415), All CpCV-E, CpCV-F (KC172700)	RMCT	1.18×10^{-11}
4	94-170	CpCV-A (KC172684, KC172685), All CpCV-E	CpCV-F (KC172700)	RGT	2.23×10^{-8}
5	1942-2543	CpCDV-A (KC172653, KC172654, KC172655, KC172663, KC172657, KC172661FR687959)	All CpCDV-F	RGBMT	1.71×10^{-07}
6	1520-1798	CpCDV-H, CpCDV-I, CpCDV-J, CpCDV-E (AM933135, AM933135)	Unknown	RMCT	2.71×10^{-07}
7	1947-205	All CpCDV-K	Unknown	RMCT	4.70×10^{-07}
8	614-1083	CpCDV-J	All CpCDV-K	RMCT	1.62×10^{-05}
9	66-110	Unknown	CpCDV-H (KC172676), CpCDV-C (AM849097)	RMC	7.23×10^{-03}
10	2543-150	Unknown	CpCDV-C (AM849097, AM900416)	RGMCT	6.64×10^{-08}
11	1259-1580	All CpCDV-D	Unknown	RGMCT	4.52×10^{-06}
12	2555-504*	CpCDV-F (KC172666, KC172669)	CpCDV-K (KC172680)	RMCT	5.36×10^{-05}
13	1257-1364	CpCDV-J, CpCDV-H	Unknown	RGB	9.09×10^{-08}
14	4-57	CpCV-C (JN989417)	Unknown, All CpCV-E	RGB	5.22×10^{-05}
15	1279-1343	CpCDV-K (KC172680)	All CpCDV-G	RGB	1.11×10^{-06}
16	1159-2487	All CpCAV	Unknown	MCST	1.90×10^{-11}
Inter-species recombination					
A	1061-1366	All CpCV-E	All TYDV-A	RGMCT	9.24×10^{-51}
B	1447-2102	All CpCV-A, All CpCV-E, All CpCAV, All CpCV-B, CpCV-F, CpCV-C (JN989416, JN989417)	All CpCDV	RGMCT	7.44×10^{-19}
C	111-159	All CpCV-E, CpCV-F, CpCV-A (KC172684, KC172683, KC172685, GU256530, JN989414, JN989415)	All TYDV-A	RGT	8.14×10^{-15}
D	1279-1335	CpCDV-K (KC172680)	All TYDV-A, CpCV-A (GU256530)	RGBM	6.16×10^{-10}
E	634-1027	Unknown	CpYV	RMC	1.64×10^{-03}
F	2078-2345	All CpCV-B	All CpCAV	RMC	1.87×10^{-06}
G	1247-1297	All CpCV-E	Unknown	RBS	5.23×10^{-05}
H	1345-1634	CpCDV-E (AM933135, AM933134), All CpCDV-A, All CpCDV-F, CpCDV-H, CpCDV-I, CpCDV-J	All CpCAV	RMS	3.35×10^{-04}
I	1725-1828	TYDV-A (JN9899443, KC172702, JN989440, JN989441, JN989444, M81103)	All CpCDV-A	RMCT	4.72×10^{-03}
J	1706-2509	CpCDV-C (AM850136, AM849097, AM900416)	All CpCAV	RBMCs	3.58×10^{-29}

RDP (R) GENCONV (G), BOOTSCAN (B), MAXCHI (M), CHIMERA (C), SISCAN (S), LARD (L) and 3SEQ (T).

* = The actual breakpoint position is undetermined.

4.4.3 The geographical origin of the dicot-infecting mastreviruses

Despite 10 years of sampling effort dicotyledonous plant species and using methods such as rolling circle amplification and next generation sequencing to identify and recover circular ssDNA viruses from infected plant material, the only regions of the world where dicot infecting mastreviruses have been conclusively identified are the Middle East, East Africa, Australia and South Africa. However, fragments of a dicot-infecting mastrevirus-like genome have been discovered through deep sequencing of small RNAs extracted from Peruvian sweet-potatoes (Kreuze *et al.*, 2009), suggesting that the currently known distribution of these viruses is almost definitely an under-estimation of their geographical range. It is nevertheless possible for us to determine which of the regions where these viruses have been sampled is nearest to their geographical origin. Our results support the prevailing notion that the degree of dicot-infecting mastrevirus diversity outside of Australia is lower than that within Australia and that the dicot-infecting viruses discovered in the former regions most likely originated either in or near Australia.

The WDV-rooted ML phylogenetic tree constructed from sequences with the tracts of recombinationally-derived sequence removed, indicated that the MRCA of these viruses (the node at the root of the tree in Fig. 4.1) is probably Australian. Also, as has been suggested in previous analyses the diversity of dicot-infecting mastreviruses in Australia is clearly far greater than that seen amongst the currently sampled African, Middle-Eastern, Turkish and Indo-Pakistani sequences.

Given that the sequences examined here were sampled over a period of only 27 years (1984–2011) it was unsurprising that our three datasets yielded only weak support for the presence of a molecular clock signal (Path-O-Gen derived correlation coefficients ranging between 0.20 and 0.25). Since this indicated that the analysed datasets could not be productively used to estimate accurate nucleotide substitution rates, it was not possible for us to accurately date any of the historical dispersal events shown by our phylogeographic analyses. Nevertheless, of the various molecular clock (strict and relaxed) and demographic (constant population size, Bayesian skyline plot) models tested the constant population size + relaxed-clock model fitted the data best.

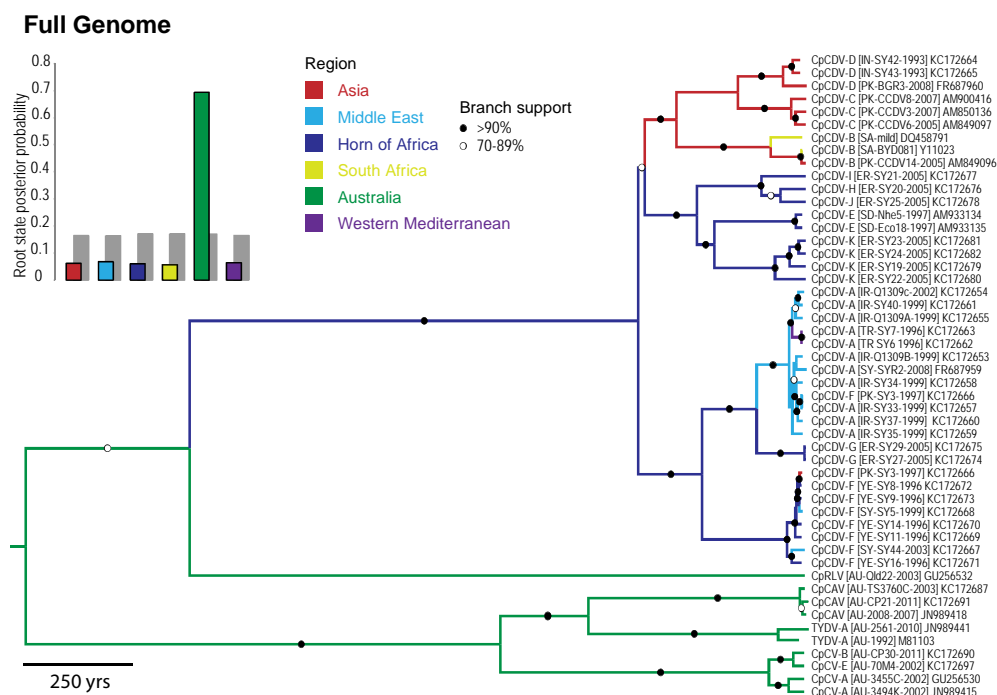
The maximum clade credibility (MCC) trees constructed using these models applied to the full genome, Rep, and CP datasets with sequences sampled from the Western Mediterranean (WM), Asian (AS) the Middle Eastern (ME), East African (EA), Southern African (SA) and Australian (AU) regions are presented in Fig. 4.3 to 4.5. For all of the analysed datasets Australia was indicated the most likely origin of the MRCA of all the analysed viruses (note the colour of the lines at the basal nodes of the trees in Fig 4.3 to 4.8). Specifically, Australia had 0.8735 posterior probability support as the root location state for the CP dataset, 0.8333 for the Rep dataset, and 0.6932 for the full genome dataset.

When the same data was analysed with the sampling locations randomized amongst the analysed sequences the most probable root locations were inferred to be either East Africa for the CP dataset ($P = 0.1789$) or the Middle east for the Rep ($P = 0.1697$) and full genome dataset ($P = 0.1701$) suggesting that our results were not inherently biased in favour of identifying Australia as the location of the MRCA (Fig 4.3 to 4.5).

4.4.4 Plausible routes of dicot infecting mastrevirus movement out of Australia

Collectively four statistically supported ($BF_{\log10} > 5.0$) virus movements between the six analysed locations were inferred from the three analysed datasets (Fig. 4.3). These involved initial movements out of Australia to both South Africa ($BF_{\log10} = 179.8, 69.0, 26.9$), and to the horn of Africa ($BF_{\log10} = 21.8, 25.7, 5.2$) with subsequent dispersal from the Middle East to Asia ($BF_{\log10} = 550.1, 56.7, 363.1$), and from horn of Africa to the Middle East ($BF_{\log10} = 203.7, 435.4, 41.9$) for the full genome, Rep and CP datasets respectively.

A.



B.

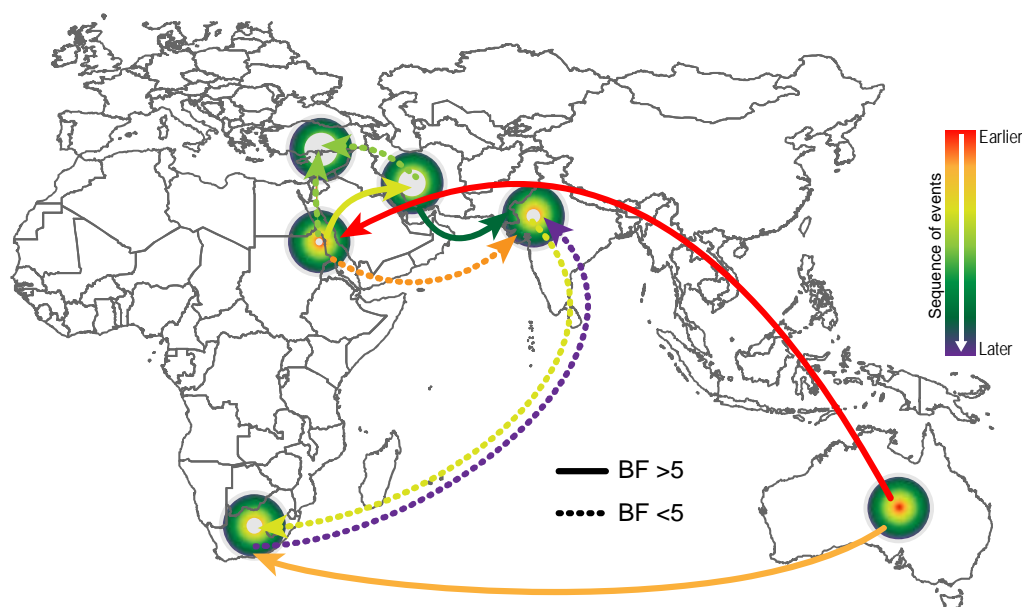


Figure 4.3: (A) Maximum clade credibility tree constructed from the dicot-infecting mastrevirus full genome dataset under the GTR + G4 nucleotide substitution model, constant population size demographic model, a relaxed-clock evolutionary model and a discretized spatial diffusion phylogeographic model. This later model considered spatial diffusion between six geographic locations and included only a randomly selected subset of 10 of the Australian mastreviruses included in Fig 1. Branches and taxon names are coloured according to the region where they were collected. Posterior support greater than 90% is indicated by a filled circle and greater than 70% by an open circle at the nodes. Probabilities obtained with randomisation of the tip locations are provided as grey bars for each location. **(B)** Plausible historical movement pathways of dicot-infecting mastreviruses inferred using the full genome dataset. The spatial dynamics of dicot-infecting mastreviruses movements were inferred using the discrete phylogeographic model considering only the six geographical regions from which the analysed viruses were sampled.

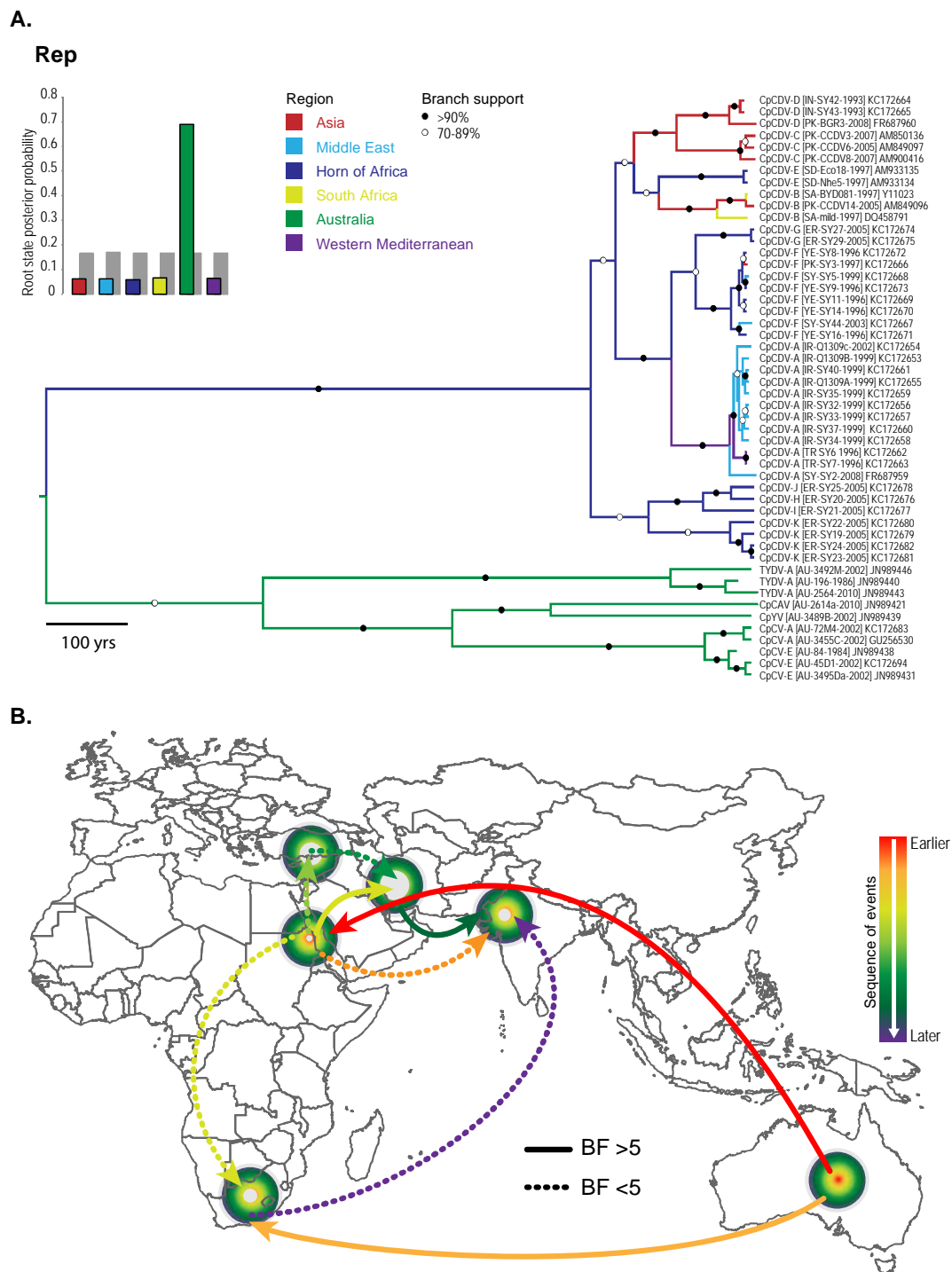


Figure 4.4: (A) Maximum clade credibility tree constructed from the dicot-infecting mastrevirus Rep dataset under the GTR + G4 nucleotide substitution model, constant population size demographic model, a relaxed-clock evolutionary model and a discretized spatial diffusion phylogeographic model. This later model considered spatial diffusion between six geographic locations and included only a randomly selected subset of 10 of the Australian mastreviruses included in Fig 1. Branches and taxon names are coloured according to the region where they were collected. Posterior support greater than 90% is indicated by a filled circle and greater than 70% by an open circle at the nodes. Probabilities obtained with randomisation of the tip locations are provided as grey bars for each location. **B.** Plausible historical movement pathways of dicot-infecting mastreviruses inferred using Rep dataset. The spatial dynamics of dicot-infecting mastreviruses movements were inferred using the discrete phylogeographic model considering only the six geographical regions from which the analysed viruses were sampled.

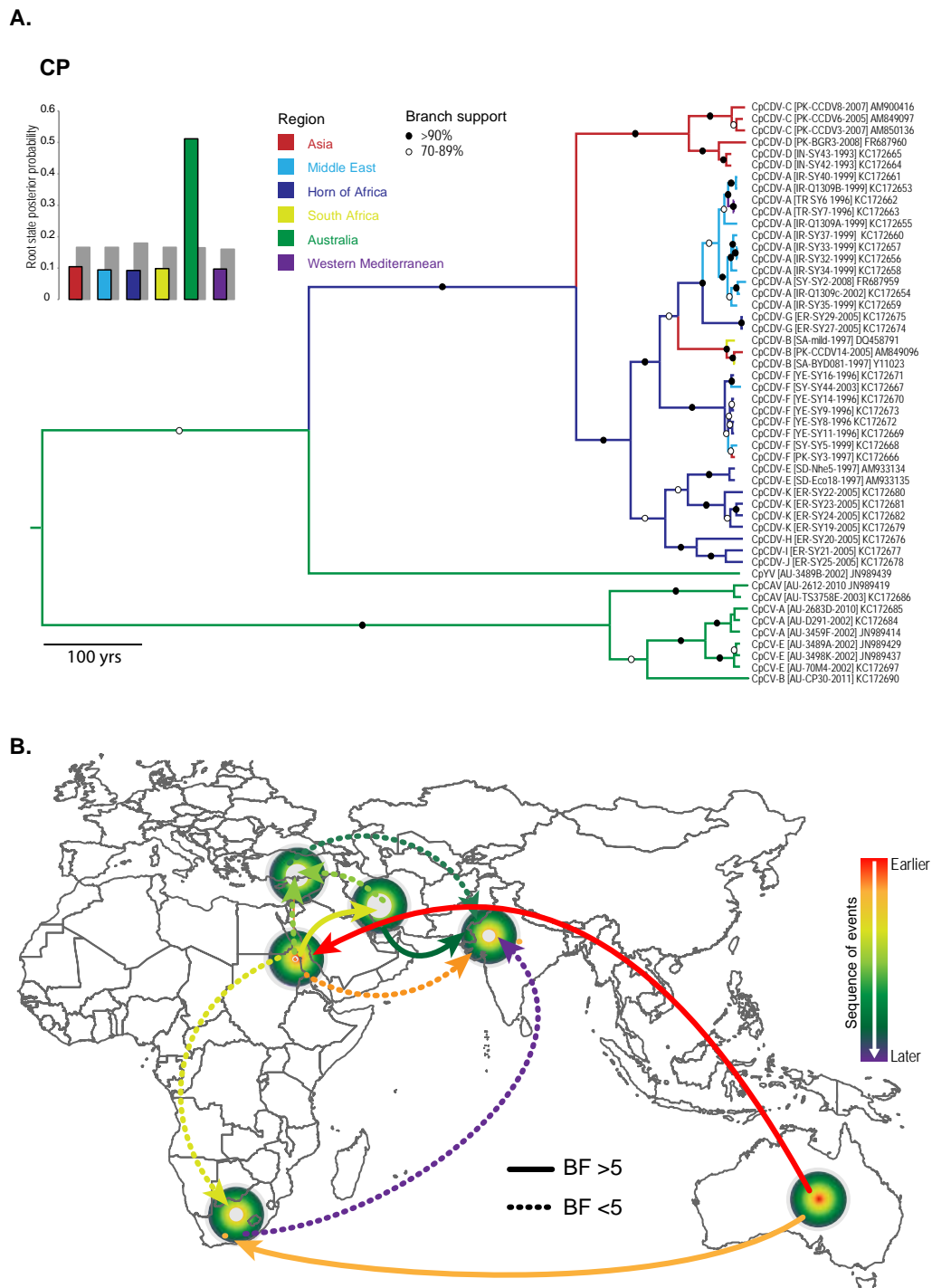


Figure 4.5: (A) Maximum clade credibility tree constructed from the dicot-infecting mastrevirus CP dataset under the GTR + G4 nucleotide substitution model, constant population size demographic model, a relaxed-clock evolutionary model and a discretized spatial diffusion phylogeographic model. This later model considered spatial diffusion between six geographic locations and included only a randomly selected subset of 10 of the Australian mastreviruses included in Fig 1. Branches and taxon names are coloured according to the region where they were collected. Posterior support greater than 90% is indicated by a filled circle and greater than 70% by an open circle at the nodes. Probabilities obtained with randomisation of the tip locations are provided as grey bars for each location. **(B)** Plausible historical movement pathways of dicot-infecting mastreviruses inferred using the CP dataset. The spatial dynamics of dicot-infecting mastreviruses movements were inferred using the discrete phylogeographic model considering only the six geographical regions from which the analysed viruses were sampled.

Full genome

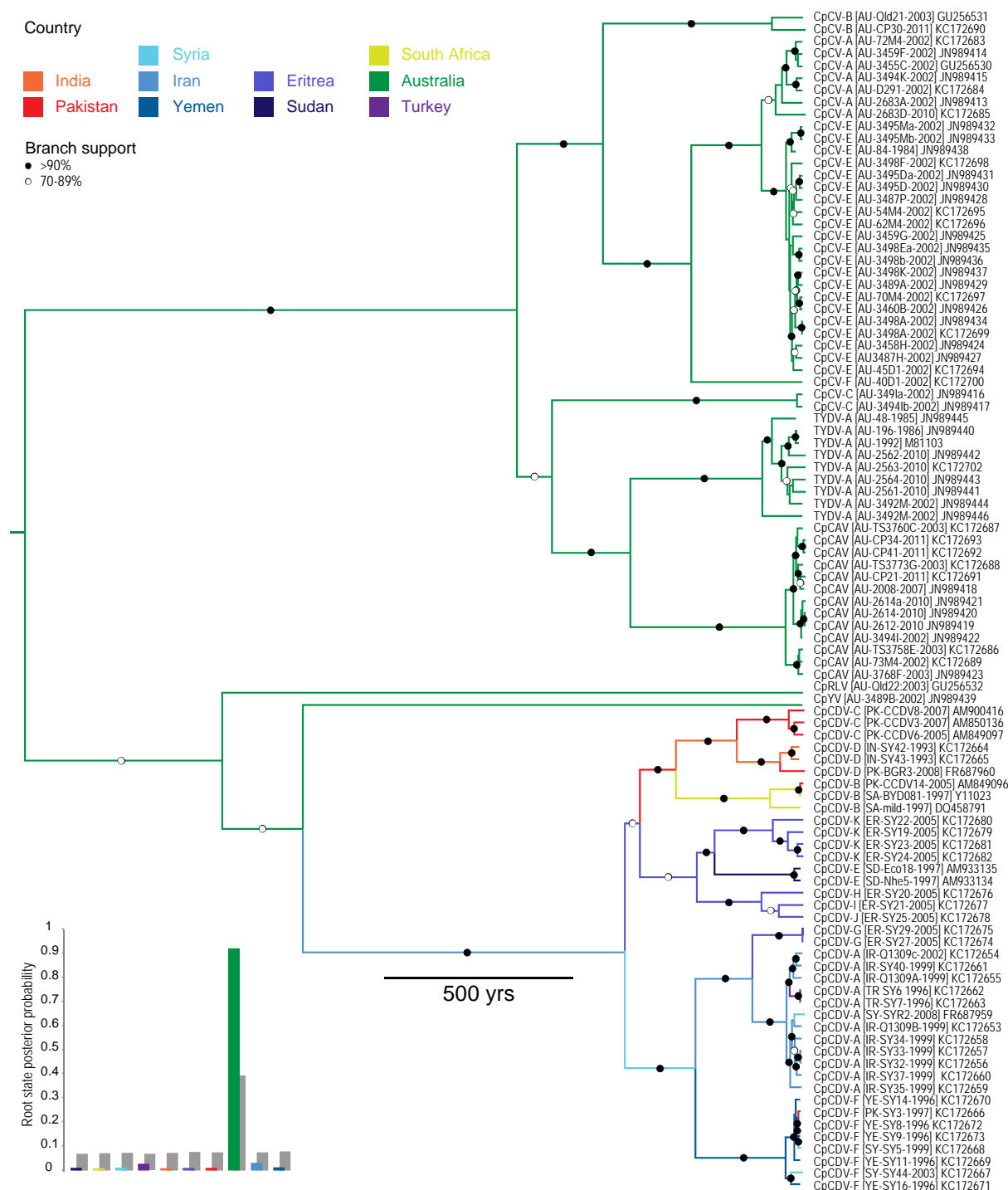


Figure 4.6: Maximum clade credibility trees for the full genome dicot-infecting mastrevirus alignments constructed under the GTR + G4 nucleotide substitution model and a constant population size relaxed-clock evolutionary model with discretized spatial diffusion. This analysis considered a model of spatial diffusion between the 10 locations. Branches and taxon names are coloured by country of collection and colour gradients on the branches represent inferred historical migrations. Posterior support greater than 90% is indicated by a filled circle and greater than 70% by an open circle at the nodes. Probabilities obtained with randomisation of the tip locations are provided as grey bars for each location.

Rep

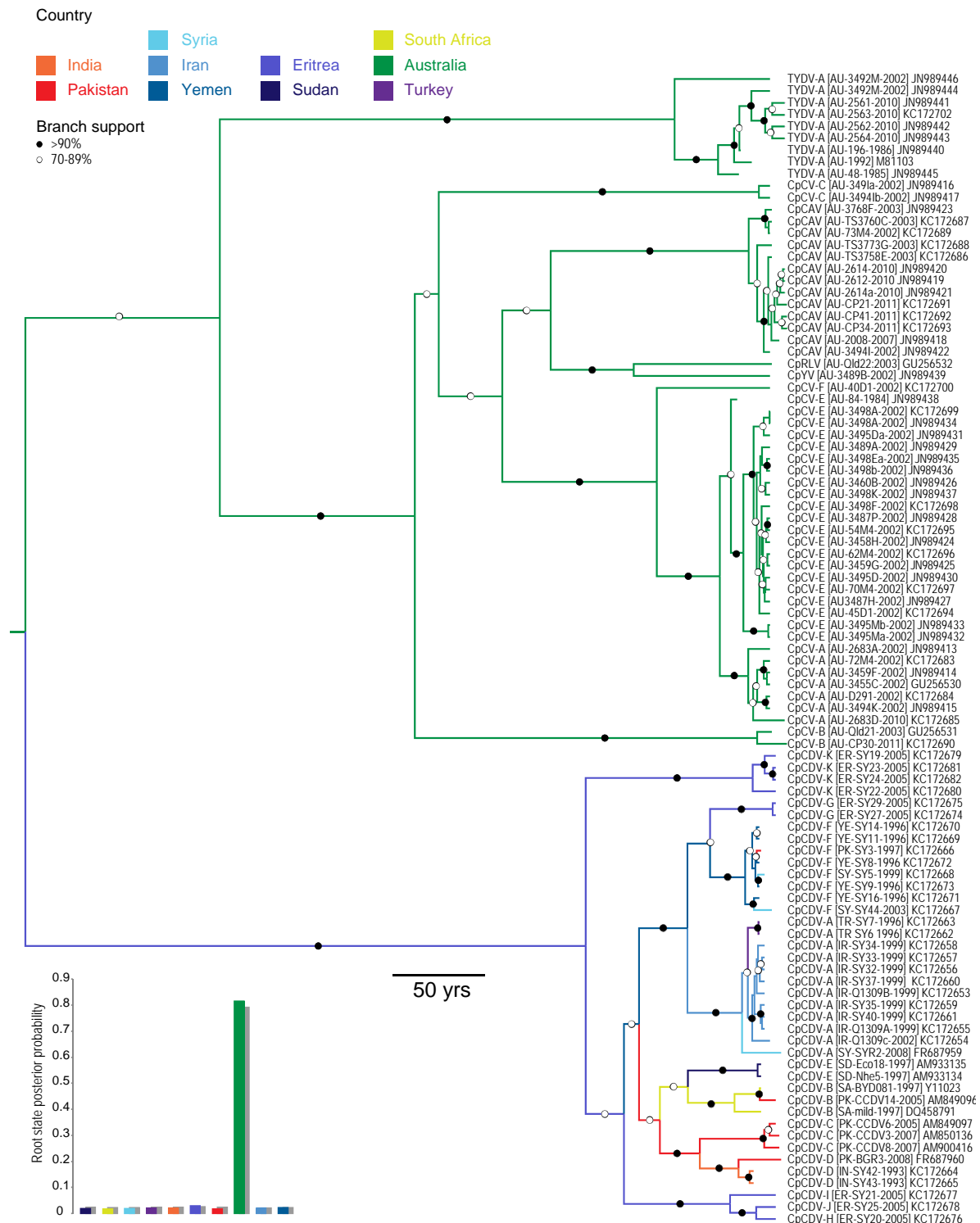


Figure 4.7: Maximum clade credibility trees for the Rep dataset of dicot-infecting mastrevirus alignments constructed under the GTR + G4 nucleotide substitution model and a constant population size relaxed-clock evolutionary model with discretized spatial diffusion. This analysis considered a model of spatial diffusion between the 10 locations. Branches and taxon names are coloured by country of collection and colour gradients on the branches represent inferred historical migrations. Posterior support greater than 90% is indicated by a filled circle and greater than 70% by an open circle at the nodes. Probabilities obtained with randomisation of the tip locations are provided as grey bars for each location.

CP

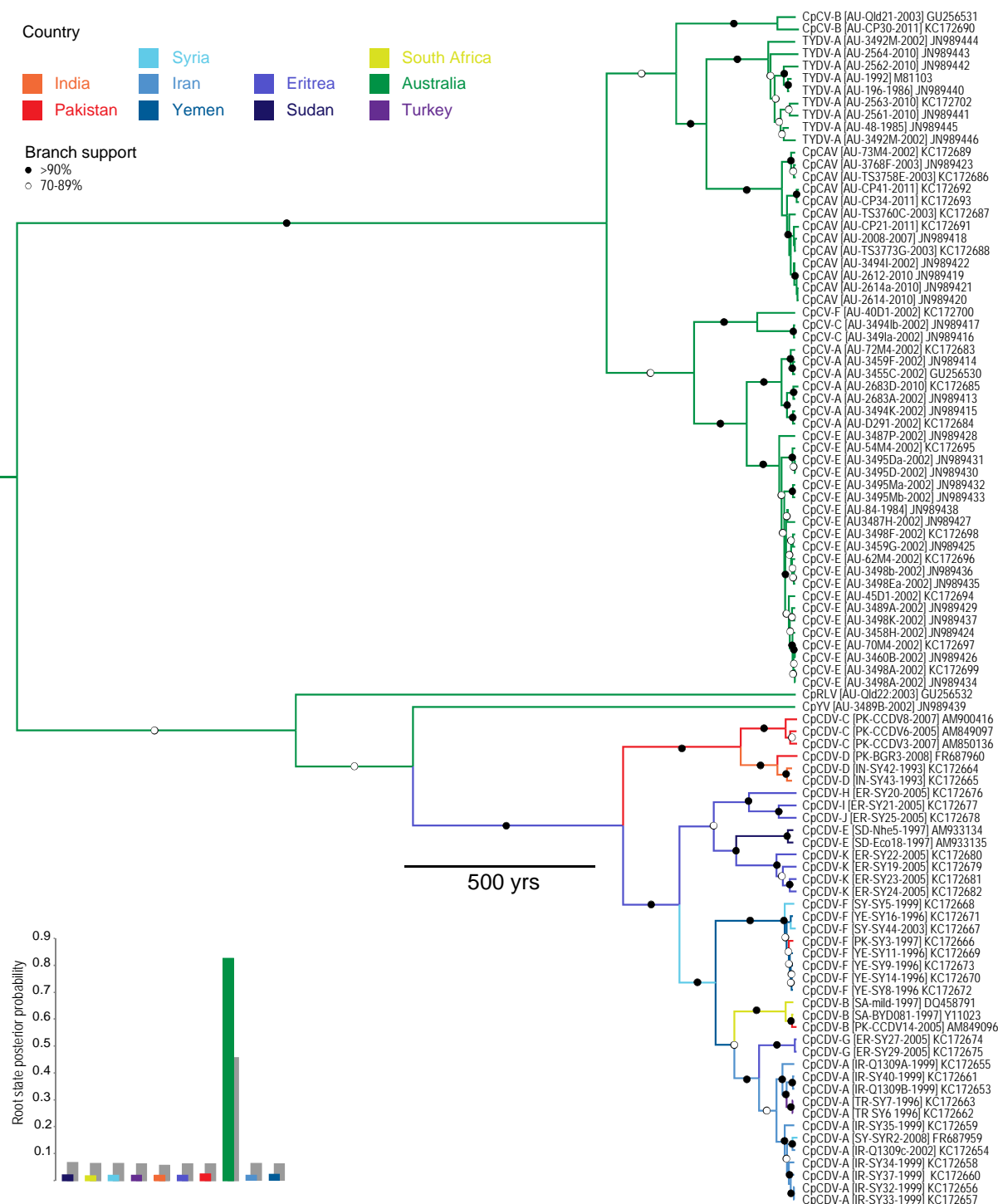


Figure 4.8: Maximum clade credibility trees for the CP dataset of dicot-infecting mastrevirus alignments constructed under the GTR + G4 nucleotide substitution model and a constant population size relaxed-clock evolutionary model with discretized spatial diffusion. This analysis considered a model of spatial diffusion between the 10 locations. Branches and taxon names are coloured by country of collection and colour gradients on the branches represent inferred historical migrations. Posterior support greater than 90% is indicated by a filled circle and greater than 70% by an open circle at the nodes. Probabilities obtained with randomisation of the tip locations are provided as grey bars for each location.

4.5 Concluding remarks

Dicot-infecting mastreviruses have been identified in Australia, Africa, the Middle East and the Indian subcontinent as potentially important crop pathogens. This study extends our current knowledge of the diversity of these viruses within these regions, with the addition of 49 full genomes. Amongst these genomes are isolates of seven new divergent strains from two different species. Of particular interest is our recombination analysis which revealed a surprisingly high level of inter-species recombination events between dicot-infecting mastrevirus from two geographically distant regions, a pattern which while consistent with that found in dicot-infecting begomoviruses and the monocot-infecting mastreviruses (Kraberger *et al.*, 2012; Shepherd *et al.*, 2010; Varsani *et al.*, 2008a; Varsani *et al.*, 2008b, Chapter 2). Such a high frequency of recombination events coupled with evidence that recombination has likely contributed to the emergence of various geminiviruses as agricultural pests during modern times (Rocha *et al.*, 2013; Varsani *et al.*, 2008b), highlights the importance of continual surveillance to monitor for the presence and identities of these viruses in the environment so as to identify potentially new pathogens that may evolve to threaten agriculture.

Pulses were among the first cultivated plants, with some of the oldest archaeobotanical evidence indicating that the Middle East is one of ancient centres of this practice (Mikić, 2012; Tanno & Willcox, 2006). Given that the Middle East and surrounding countries have such a long history of the cultivation of pulses in comparison with Australia it is surprising that Australia harbours a greater diversity of dicot-infecting mastreviruses than Africa, the Middle East and Indian Subcontinent combined. The corrective measures that we have taken to account for recombination and sampling biases, strengthen our conclusion that the MRCA of the currently known dicot-infecting mastreviruses is most likely nearer Australia than the other sampling locations that were considered. It is nevertheless important to stress that Australia is merely the region amongst those that have been sampled where the MRCA of the analysed sequences originated. The MRCA of these sequences could have actually existed in any of the many regions of the world where samples have not been collected, with descendants of these sequences having simply passed through Australia *en route* to the other geographical regions that have been considered here. Similarly, the MRCA of the sequences considered here is not necessarily the MRCA of all the dicot-infecting mastreviruses

currently circulating on Earth, and is almost certainly also not the “first” mastrevirus that infected dicotyledonous hosts. Given that fragments of a highly divergent virus genome resembling those of dicot-infecting mastreviruses has been detected in the Peruvian sweet potato germplasm collection (Kreuze *et al.*, 2009), it is entirely plausible that the viruses considered here are part of a much more diverse, but currently undiscovered, global dicot-infecting mastrevirus population.

Without much more intensive sampling of dicot-infecting mastreviruses, both in the regions considered here and across the vast areas of Asia, Africa, the Pacific Rim and the Americas where these viruses have remained unsampled, we cannot yet hope to pinpoint the actual geographical origins of either the MRCA of all dicot-infecting mastreviruses, or the location of the first dicot-infecting mastrevirus. With the application of modern molecular tools and new metagenomic approaches to mastrevirus discovery (Rosario *et al.*, 2013), we anticipate that there will be a rapid increase in the diversity of known dicot-infecting mastreviruses that should greatly increase the resolution with which the movement pathways and geographically origins of these viruses can be determined.

GenBank accession numbers: KC172653 – KC172702

Additional Table 4.1: Sampling information for the dicot-infecting mastrevirus genomes sequences determined in this study.

Sample#	Isolate	GenBank accession	Field specimen ID	Lab ID	Host	Sampling Year	Origin
1	CpCDV-A IR-Q1309A-1999	KC172655	IC149-99	Q1309A	Chickpea	1999	Iran
2	CpCDV-A IR-Q1309B-1999	KC172653	SP175-99	Q1309B	Chickpea	1999	Iran
3	CpCDV-A IR-Q1309C-2002	KC172654	IC 1-02	Q1309C	Chickpea	2002	Iran
4	CpCDV-A IR-SY32-1999	KC172656	IC 140-99	SY32	Chickpea	1999	Iran
5	CpCDV-A IR-SY33-1999	KC172657	IC 143-99	SY33	Chickpea	1999	Iran
6	CpCDV-A IR-SY34-1999	KC172658	IC 147-99	SY34	Chickpea	1999	Iran
7	CpCDV-A IR-SY35-1999	KC172659	IC 148-99	SY35	Chickpea	1999	Iran
8	CpCDV-A IR-SY37-1999	KC172660	IC 152-99	SY37	Chickpea	1999	Iran
9	CpCDV-A IR-SY40-1999	KC172661	IP 175-99	SY40	Field pea	1999	Iran
10	CpCDV-A TR-SY6-1996	KC172662	TC41-96	SY06	Chickpea	1996	Turkey
11	CpCDV-A TR-SY7-1996	KC172663	TC40-96	SY07	Chickpea	1996	Turkey
12	CpCDV-D IN-SY42-1993	KC172664	ICRISAT 1	SY42	Chickpea	1993	India
13	CpCDV-D IN-SY43-1993	KC172665	ICRISAT 2	SY43	Field pea	1993	India
14	CpCDV-F PK-SY3-1997	KC172666	PL148-97	SY03	Lentil	1997	Pakistan
15	CpCDV-F SY-SY44-2003	KC172667	SC 3-03	SY44	Chickpea	2003	Syria
16	CpCDV-F SY-SY5-1999	KC172668	SC54-99	SY05	Chickpea	1999	Syria
17	CpCDV-F YE-SY11-1996	KC172669	YeL8-96	SY11	Lentil	1996	Yemen
18	CpCDV-F YE-SY14-1996	KC172670	YeV18-96	SY14	Faba bean	1996	Yemen
19	CpCDV-F YE-SY16-1996	KC172671	YeV26-96	SY16	Faba bean	1996	Yemen
20	CpCDV-F YE-SY8-1996	KC172672	YeL5-96	SY08	Lentil	1996	Yemen
21	CpCDV-F YE-SY9-1996	KC172673	YeL6-96	SY09	Lentil	1996	Yemen
22	CpCDV-G ER-SY27-2005	KC172674	ErC472-05	SY27	Chickpea	2005	Eritrea
23	CpCDV-G ER-SY29-2005	KC172675	ErC505-05	SY29	Chickpea	2005	Eritrea
24	CpCDV-H ER-SY20-2005	KC172676	ErC177-05	SY20	Chickpea	2005	Eritrea
25	CpCDV-I ER-SY21-2005	KC172677	ErC180-05	SY21	Chickpea	2005	Eritrea
26	CpCDV-J ER-SY25-2005	KC172678	ErC351-05	SY25	Chickpea	2005	Eritrea
27	CpCDV-K ER-SY19-2005	KC172679	ErC86-05	SY19	Chickpea	2005	Eritrea
28	CpCDV-K ER-SY22-2005	KC172680	ErC243-05	SY22	Chickpea	2005	Eritrea
29	CpCDV-K ER-SY23-2005	KC172681	ErC345-05	SY23	Chickpea	2005	Eritrea
30	CpCDV-K ER-SY24-2005	KC172682	ErC347-05	SY24	Chickpea	2005	Eritrea
31	CpCV-A AU-2683D-2010	KC172685	2683D	2683D	Chickpea	2010	Australia
32	CpCV-A AU-72M4-2002	KC172683	3460D	72M4	Chickpea	2002	Australia
33	CpCV-A AU-D291-2002	KC172684	3494K	D291	Chickpea	2002	Australia
35	CpCV-B AU-CP30-2011	KC172690	3246	CP30	Chickpea	2011	Australia
34	CpCAV AU-CP21-2011	KC172691	3245	CP21	Chickpea	2011	Australia
37	CpCAV AU-CP34-2011	KC172693	3247	CP34	Chickpea	2011	Australia
36	CpCAV AU-CP41-2011	KC172692	3248	CP41	Chickpea	2011	Australia
38	CpCAV AU-73M4-2002	KC172689	3460E	73M4	Chickpea	2002	Australia
39	CpCAV AU-TS3758E-2003	KC172686	3758E	TS3758E	Chickpea	2003	Australia
40	CpCAV AU-TS3768C-2003	KC172687	3768C	TS3760C	Chickpea	2003	Australia
41	CpCAV AU-TS3773G-2003	KC172688	3773G	TS3773G	Chickpea	2003	Australia

Sample #	Isolate	GenBank accession	Field specimen ID	Lab ID	Host	Sampling Year	Origin
42	CpCV-E AU-45D1-2002	KC172694	3487O	45D1	Chickpea	2002	Australia
43	CpCV-E AU-54M4-2002	KC172695	3459D	54M4	Chickpea	2002	Australia
44	CpCV-E AU-62M4-2002	KC172696	3459L	62M4	Chickpea	2002	Australia
45	CpCV-E AU-70M4-2002	KC172697	3460B	70M4	Chickpea	2002	Australia
46	CpCV-E AU-3498A-2002	KC172699	3498A	3498A	Chickpea	2002	Australia
47	CpCV-E AU-3498F-2002	KC172698	3498F	3498F	Chickpea	2002	Australia
48	CpCV-F AU-40D1-2002	KC172700	3487J	40D1	Chickpea	2002	Australia
49	TYDV-A AU-2563-2010	KC172702	2563	2563	Bean	2010	Australia

4.6 References

- Ali, M. A., Kumari, S. G., Makkouk, K. H. & Hassan, M. M. (2004). Chickpea chlorotic dwarf virus, CpCDV naturally infects Phaseolus bean and other wild species in the Gezira region of Sudan. *Arab Journal of Plant Protection* **22**, 96.
- Bielejec, F., Rambaut, A., Suchard, M. A. & Lemey, P. (2011). SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. *Bioinformatics* **27**, 2910-2912.
- Boni, M. F., Posada, D. & Feldman, M. W. (2007). An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176**, 1035-1047.
- De Bruyn, A., Villemot, J., Lefeuvre, P., Villar, E., Hoareau, M., Harimalala, M., Abdoul-Karime, A. L., Abdou-Chakour, C., Reynaud, B., Harkins, G. W., Varsani, A., Martin, D. P. & Lett, J. M. (2012). East African cassava mosaic-like viruses from Africa to Indian ocean islands: molecular diversity, evolutionary history and geographical dissemination of a bipartite begomovirus. *BMC Evolutionary Biology* **12**, 228.
- Dekker, E. L., Woolston, C. J., Xue, Y., Cox, B. & Mullineaux, P. M. (1991). Transcript mapping reveals different expression strategies for the bicistronic RNAs of the geminivirus wheat dwarf virus. *Nucleic Acids Research* **19**, 4075-4081.
- Drummond, A., Pybus, O. G. & Rambaut, A. (2003). Inference of viral evolutionary rates from molecular sequences. *Advances in Parasitology* **54**, 331-358.
- Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**, 1969-1973.
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797.
- Farzadfar, S., Pourrahim, R., Golnaraghi, A. R., Shahraeen, N. & Makkouk, K. M. (2002). First report of sugar beet and bean as natural hosts of Chickpea chlorotic dwarf virus in Iran. *Plant Pathology* **51**, 795-795.
- Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. (2000). Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573-582.
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**, 307-321.
- Hadfield, J., Thomas, J. E., Schwinghamer, M. W., Krabberger, S., Stainton, D., Dayaram, A., Parry, J. N., Pande, D., Martin, D. P. & Varsani, A. (2012). Molecular characterisation of dicot-infecting mastreviruses from Australia. *Virus Research* **166**, 13-22.
- Halley-Stott, R. P., Tanzer, F., Martin, D. P. & Rybicki, E. P. (2007). The complete nucleotide sequence of a mild strain of Bean yellow dwarf virus. *Arch Virol* **152**, 1237-1240.
- Harrison, B. D. (1985). Advances in Geminivirus Research. *Annual Review of Phytopathology* **23**, 55-82.

- Heyraud, F., Matzeit, V., Schaefer, S., Schell, J. & Gronenborn, B. (1993).** The conserved nonanucleotide motif of the geminivirus stem-loop sequence promotes replicational release of virus molecules from redundant copies. *Biochimie* **75**, 605-615.
- Horn, N. M., Reddy, S. V. & Reddy, D. V. R. (1994).** Virus-vector relationships of chickpea chlorotic dwarf geminivirus and the leafhopper *Orosius orientalis* (Hemiptera: Cicadellidae). *Annals of Applied Biology* **124**, 441-450.
- Horn, N. M., Reddy, S. V., Roberts, I. M. & Reddy, D. V. R. (1993).** Chickpea chlorotic dwarf virus, a new leafhopper-transmitted geminivirus of chickpea in India. *Annals of Applied Biology* **122**, 467-479.
- Jeske, H., Lütgemeier, M. & Preiß, W. (2001).** DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. *The EMBO journal* **20**, 6158-6167.
- Knights, E. J., Açıkgöz, N., Warkentin, T., Bejiga, G., Yadav, S. S. & Sandhu, J. S. (2007).** Area, production, and distribution. *Chickpea breeding and management*, 167-178.
- Kraberger, S., Thomas, J. E., Geering, A. D. W., Dayaram, A., Stainton, D., Hadfield, J., Walters, M., Parmenter, K. S., van Brunschot, S., Collings, D. A., Martin, D. P. & Varsani, A. (2012).** Australian monocot-infecting mastrevirus diversity rivals that in Africa. *Virus Research* **169**, 127-136.
- Kreuze, J. F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I. & Simon, R. (2009).** Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. *Virology* **388**, 1-7.
- Kumari, S. G., Makkouk, K. M., Attar, N., Ghulam, W. & Lesemann, D. E. (2004).** First Report of Chickpea chlorotic dwarf virus infecting spring chickpea in Syria. *Plant Disease* **88**, 424-424.
- Kumari, S. G., Makkouk, K. M., Loh, M. H., Negassi, K., Tsegay, S., Kidane, R., Kibret, A. & Tesfatsion, Y. (2008).** Viral diseases affecting chickpea crops in Eritrea. *Phytopathologia Mediterranea* **47**, 42-49.
- Lefeuvre, P., Martin, D. P., Harkins, G., Lemey, P., Gray, A. J., Meredith, S., Lakay, F., Monjane, A., Lett, J. M., Varsani, A. & Heydarnejad, J. (2010).** The spread of tomato yellow leaf curl virus from the Middle East to the world. *PLoS pathogens* **6**, e1001164.
- Lefeuvre, P., Martin, D. P., Hoareau, M., Naze, F., Delatte, H., Thierry, M., Varsani, A., Becker, N., Reynaud, B. & Lett, J.-M. (2007).** Begomovirus ‘melting pot’ in the south-west Indian Ocean islands: molecular diversity and evolution through recombination. *Journal of General Virology* **88**, 3458-3468.
- Makkouk, K. M., Rizkallah, L., Kumari, S. G., Zaki, M. & Enein, R. A. (2003).** First record of Chickpea chlorotic dwarf virus (CpCDV) affecting faba bean (*Vicia faba*) crops in Egypt. *Plant Pathology* **52**, 413-413.
- Martin, D. P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P. & Varsani, A. (2011a).** Recombination in Eukaryotic Single Stranded DNA Viruses. *Viruses* **3**, 1699-1738.

- Martin, D. P., Briddon, R. W. & Varsani, A. (2011b).** Recombination patterns in dicot-infecting mastreviruses mirror those found in monocot-infecting mastreviruses. *Arch Virol* **156**, 1463-1469.
- Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D. & Lefevre, P. (2010).** RDP3: A flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**, 2462-2463.
- Martin, D. P., Posada, D., Crandall, K. A. & Williamson, C. (2005).** A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Research and Human Retroviruses* **21**, 98-102.
- Martin, D. P., Willment, J. A., Billharz, R., Velders, R., Odhiambo, B., Njuguna, J., James, D. & Rybicki, E. P. (2001).** Sequence diversity and virulence in *Zea mays* of Maize streak virus isolates. *Virology* **288**, 247-255.
- Mikić, A. (2012).** Origin of the Words Denoting Some of the Most Ancient Old World Pulse Crops and Their Diversity in Modern European Languages. *PLoS ONE* **7**, e44512.
- Monjane, A. L., Harkins, G. W., Martin, D. P., Lemey, P., Lefevre, P., Shepherd, D. N., Oluwafemi, S., Simuyandi, M., Zinga, I., Komba, E. K., Lakoutene, D. P., Mandakombo, N., Mboukoulida, J., Semballa, S., Tagne, A., Tiendrébéogo, F., Erdmann, J. B., van Antwerpen, T., Owor, B. E., Flett, B., Ramusi, M., Windram, O. P., Syed, R., Lett, J.-M., Briddon, R. W., Markham, P. G., Rybicki, E. P. & Varsani, A. (2011).** Reconstructing the history of Maize streak virus Strain A dispersal to reveal diversification hot spots and its origin in Southern Africa. *Journal of Virology* **85**, 9623-9636.
- Morris, B. A. M., Richardson, K. A., Haley, A., Zhan, X. & Thomas, J. E. (1992).** The nucleotide sequence of the infectious cloned dna component of tobacco yellow dwarf virus reveals features of geminiviruses infecting monocotyledonous plants. *Virology* **187**, 633-642.
- Muhire, B., Martin, D., Brown, J., Navas-Castillo, J., Moriones, E., Zerbini, F. M., Rivera-Bustamante, R., Malathi, V. G., Briddon, R. & Varsani, A. (2013).** A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). *Archives of Virology* **158**, 1411-1424.
- Mullineaux, P. M., Guerineau, F. & Accotto, G.-P. (1990).** Processing of complementary sense RNAs of Digitaria streak virus in its host and in transgenic tobacco. *Nucleic Acids Research* **18**, 7259-7265.
- Mumtaz, H., Kumari, S., Mansoor, S., Martin, D. & Briddon, R. (2011).** Analysis of the sequence of a dicot-infecting mastrevirus (family *Geminiviridae*) originating from Syria. *Virus Genes* **42**, 422-428.
- Nahid, N., Amin, I., Mansoor, S., Rybicki, E., van der Walt, E. & Briddon, R. (2008).** Two dicot-infecting mastreviruses (family *Geminiviridae*) occur in Pakistan. *Arch Virol* **153**, 1441-1451.
- Owor, B. E., Shepherd, D. N., Taylor, N. J., Edema, R., Monjane, A. L., Thomson, J. A., Martin, D. P. & Varsani, A. (2007).** Successful application of FTA((R)) Classic Card technology and use of bacteriophage phi 29 DNA polymerase for large-scale field sampling and cloning of complete maize streak virus genomes. *Journal of Virological Methods* **140**, 100-105.

- Padidam, M., Sawyer, S. & Fauquet, C. M. (1999).** Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218-225.
- Posada, D. & Crandall, K. A. (1998).** MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817-818.
- Rocha, C. S., Castillo-Urquiza, G. P., Lima, A. T., Silva, F. N., Xavier, C. A., Hora-Junior, B. T., Beserra-Junior, J. E., Malta, A. W., Martin, D. P., Varsani, A., Alfenas-Zerbini, P., Mizubuti, E. S. & Zerbini, F. M. (2013).** Brazilian begomovirus populations are highly recombinant, rapidly evolving, and segregated based on geographical location. *Journal of Virology* **87**, 5784–5799.
- Rosario, K., Padilla-Rodriguez, M., Kraberg, S., Stainton, D., Martin, D. P., Breitbart, M. & Varsani, A. (2013).** Discovery of a novel mastrevirus and alphasatellite-like circular DNA in dragonflies (Ephemeroptera) from Puerto Rico. *Virus Research* **171**, 231-237.
- Schalk, H. J., Matzeit, V., Schiller, B., Schell, J. & Gronenborn, B. (1989).** Wheat dwarf virus, a geminivirus of graminaceous plants needs splicing for replication. *EMBO Journal* **8**, 359-364.
- Schwinghamer, M., Thomas, J., Schilg, M., Parry, J., Dann, E., Moore, K. & Kumari, S. (2010).** Mastreviruses in chickpea (*Cicer arietinum*) and other dicotyledonous crops and weeds in Queensland and northern New South Wales, Australia. *Australasian Plant Pathology* **39**, 551-561.
- Shepherd, D. N., Martin, D. P., Lefevre, P., Monjane, A. L., Owor, B. E., Rybicki, E. P. & Varsani, A. (2008).** A protocol for the rapid isolation of full geminivirus genomes from dried plant tissue. *Journal of Virological Methods* **149**, 97-102.
- Shepherd, D. N., Martin, D. P., Van Der Walt, E., Dent, K., Varsani, A. & Rybicki, E. P. (2010).** Maize streak virus: An old and complex 'emerging' pathogen. *Molecular Plant Pathology* **11**, 1-12.
- Smith, J. M. (1992).** Analyzing the mosaic structure of genes. *Journal of Molecular Evolution* **34**, 126-129.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. (2011).** MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**, 2713-2739.
- Tanno, K. I. & Willcox, G. (2006).** The origins of cultivation of *Cicer arietinum* L. and *Vicia faba* L.: Early finds from Tell el-Kerkh, north-west Syria, late 10th millennium B.P. *Vegetation History and Archaeobotany* **15**, 197-204.
- Thomas, J., Parry, J., Schwinghamer, M. & Dann, E. (2010).** Two novel mastreviruses from chickpea (*Cicer arietinum*) in Australia. *Arch Virol* **155**, 1777-1788.
- Varsani, A., Monjane, A. L., Donaldson, L., Oluwafemi, S., Zinga, I., Komba, E. K., Plakoutene, D., Mandakombo, N., Mboukoulida, J., Semballa, S., Briddon, R. W., Markham, P. G., Lett, J. M., Lefevre, P., Rybicki, E. P. & Martin, D. P. (2009a).** Comparative analysis of *Panicum* streak virus and Maize streak virus diversity, recombination patterns and phylogeography. *Virology Journal* **6**, e194.

- Varsani, A., Oluwafemi, S., Windram, O., Shepherd, D., Monjane, A., Owor, B., Rybicki, E., Lefeuvre, P. & Martin, D. (2008a).** Panicum streak virus diversity is similar to that observed for maize streak virus. *Arch Virol* **153**, 601-604.
- Varsani, A., Shepherd, D. N., Dent, K., Monjane, A. L., Rybicki, E. P. & Martin, D. P. (2009b).** A highly divergent South African geminivirus species illuminates the ancient evolutionary history of this family. *Virology Journal* **6**, e36.
- Varsani, A., Shepherd, D. N., Monjane, A. L., Owor, B. E., Erdmann, J. B., Rybicki, E. P., Peterschmitt, M., Briddon, R. W., Markham, P. G., Oluwafemi, S., Windram, O. P., Lefeuvre, P., Lett, J. M. & Martin, D. P. (2008b).** Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. *Journal of General Virology* **89**, 2063-2074.
- Wright, E. A., Heckel, T., Groenendijk, J., Davies, J. W. & Boulton, M. I. (1997).** Splicing features in maize streak virus virion- and complementary-sense gene expression. *The Plant Journal* **12**, 1285-1297.

Chapter 5

Identification of an Australian-like dicot-infecting mastrevirus in Pakistan

Contents

5.1	Abstract.....	186
5.2	Introduction.....	187
5.3	Materials and methods.....	188
5.3.1	DNA extraction and isolation of mastrevirus genomes.....	188
5.3.2	Pairwise identity comparisons and construction of phylogenetic trees.....	188
5.4	Results and discussion	189
5.4.1	Classification of CpCDV genomes.....	189
5.4.2	Discovery of two Novel Australian-like mastrevirus isolates	192
5.5	Concluding remarks.....	197
5.6	References.....	198

This body of work has been published in Archives of Virology and is presented in a similar manner to that of the publication:

Kraberger, S., Mumtaz, H., Claverie, S., Martin, D. P., Briddon, R. W., Varsani, A. (2014)
Identification of an Australian-like dicot-infecting mastrevirus in Pakistan. Archives of
Virology 160, 825-830.

5.1 Abstract

Five distinct viral species in the genus *Mastrevirus* (family *Geminiviridae*) infect dicotyledonous plants in Australia, in the remainder of the world only a single dicot-infecting mastrevirus species has ever been identified. This species, *Chickpea chlorotic dwarf virus* (CpCDV), has been found infecting leguminous hosts in Africa, the Middle East and the Indian subcontinent. To further explore the diversity of CpCDV in Pakistan, ten full mastrevirus genome sequences were determined from chickpea and lentil plants. Eight of these genomes are from previously described CpCDV strains and included the first report of strain D and H isolates in Pakistan. Two other genomes derived from infected chickpea plants, are more closely related to dicot-infecting mastrevirus species found in Australia than they are to CpCDV. These two divergent genomes share less than 75% genome-wide nucleotide sequence identity with other characterised mastreviruses and therefore likely represent a second dicot-infecting mastrevirus species outside of Australia. We propose naming this species Chickpea yellow dwarf virus (CpYDV). We discuss how the presence of CpYDV in Pakistan weakens the hypothesis that Australia is the geographical origin of the dicot-infecting mastreviruses.

5.2 Introduction

An overview of the dicot-infecting mastreviruses and their potential origins is discussed in Chapter Four. Following on from Chapter 4 the motivation behind this next study was to look more closely at the dynamics of dicot-infecting mastreviruses present in a major pulse growing region. Pakistan is among the top chickpea producing regions of the world (FAOSTAT, 2013) and previously studies have reported CpCDV in Pakistan (Manzoor *et al.*, 2014; Nahid *et al.*, 2008), therefore further studies in this region may shed some light on the diversity of dicot-infecting mastreviruses circulating in this region.

Six species of dicot-infecting mastreviruses are known, five of which have only ever been identified in Australia: *Chickpea chlorosis virus* (CpCV) (Hadfield *et al.*, 2012; Krabberger *et al.*, 2013; Thomas *et al.*, 2010), *Chickpea chlorosis Australia virus* (CpCAV) (Hadfield *et al.*, 2012), *Chickpea yellows virus* (CpYV) (Hadfield *et al.*, 2012), *Chickpea redleaf virus* (CpRLV) (Thomas *et al.*, 2010) and *Tobacco yellow dwarf virus* (TYDV) (Hadfield *et al.*, 2012; Morris *et al.*, 1992; Thomas *et al.*, 2010). All dicot-infecting mastreviruses that have so far been found outside Australia all belong to a single species, *Chickpea chlorotic dwarf virus* (CpCDV) (Muhire *et al.*, 2013), the geographical range of which includes Africa, the Middle East and the Indian subcontinent.

In many of these areas CpCDV is an important biotic constraint to chickpea production (Hamed & Makkouk, 2002; Kanakala *et al.*, 2013). It appears to have a broad host-range which includes legumes such as chickpeas, lentils and common beans (Krabberger *et al.*, 2013; Liu *et al.*, 1997; Mumtaz *et al.*, 2011) peppers (Akhtar *et al.*, 2013), cotton (Manzoor *et al.*, 2013), sugar beet (Farzadfar *et al.*, 2008) and the legume weed species *Sesbenia bispinosa* (Nahid *et al.*, 2008). Between 2008 and 2012 leaves of chickpea plants showing yellowing or reddening symptoms (n=117), lentil plants with yellowing symptoms (n=27) and a common vetch (*Vivica sativa*; family *Fabaceae*) plant (n=1) with yellowing symptoms were collected. These samples were collected in areas around the cities of Bahawalnagar (n=4), Bhakkar (n=9), Chakwal (n=6), Faisalabad (n=53), Hyderabad (n=14), Jhang (n=18), Khushab (n=25), Mianwali (n=1), Muzaffargarh (n=4) and Rahim Yar Kahn (n=11) in Pakistan, to determine the incidence and diversity of CpCDV.

5.3 Materials and methods

5.3.1 DNA extraction and isolation of mastrevirus genomes

Total genomic DNA was extracted from the leaf material of individual plant samples using the GF-1 nucleic acid extraction kit (Vivantis Technologies, Malaysia) and circular DNA was enriched using an Illustra TempliPhi Amplification Kit (GE Healthcare, USA). The resulting concatameric DNA was used as a template for the recovery of complete genomes by PCR using Kapa HiFi DNA polymerase (Kapa Biosystems, USA) and the abutting primer pair (dicot forward 5'-GAN TTG GTC CGC AGT GTA GA-3'/dicot reverse 5'-GTA CCG GWA AGA CMW CYT GG-3') designed in a conserved sequence region in the genomes of dicot-infecting mastreviruses. The PCR amplification protocol consisted of 94°C for 3 min followed by 25 cycles of 98°C for 3 min, 52°C for 30 sec and 72°C for 2 min 45 sec, followed by a final extension of 72°C for 3 min. The resulting amplicons were ligated into the plasmid pJET1.2 (Fermentas, USA) and clones were Sanger sequenced at Macrogen Inc. (Korea) by primer walking. Sequences were assembled using DNA Baser Sequence Assembler V4 (Heracle Biosoft, Romania). One of the amplicons recovered was not CpCDV and a BLASTx comparison (Altschul *et al.*, 1990) showed this amplicon only shared low identity to other mastrevirus species from Australia. To confirm this genome and screen all samples for the presence of this virus we designed specific back-to-back primers (PK37 mastre F 5'-GGT TTC TGA AGT CAC CTC TGG TG-3' and PK37 mastre R 5'-ATC GAG TCA GCC CAA CCA AAT CTG-3').

5.3.2 Pairwise identity comparisons and construction of phylogenetic trees

The complete mastrevirus genomes recovered here, together with representative genomes of each dicot-infecting mastrevirus species and strain, were managed using MEGA 5.2 (Tamura *et al.*, 2011) and aligned using MUSCLE (Edgar, 2004). Open reading frames were identified using DNAMAN V7 (Lynnon Biosoft, Canada). Maximum-likelihood phylogenetic trees were constructed for the full genome dataset, as well as the CP and Rep amino acid datasets using PHYML (Guindon *et al.*, 2010) (with the best-fit model HKY and LG, respectively, determined using jModelTest (Posada, 2009) and ProtTest (Darriba *et al.*, 2011). Phylogenetic trees were rooted with *Wheat dwarf virus* (WDV)/ *Oat dwarf virus* (ODV).

Pairwise identity comparisons of genome-wide nucleotide, CP amino acid and Rep amino acid sequences were performed using SDT v1.2 (Muhire *et al.*, 2013).

5.4 Results and discussion

5.4.1 Classification of CpCDV genomes

A total of eight CpCDV genomes were recovered from 145 symptomatic leaf samples (characteristic of that seen in mastrevirus infected pulse plants) in Pakistan between 2008 and 2012, six from chickpea samples and two from lentil samples (Table 5.1). In addition, in two chickpeas samples (PK37 and PK103) a divergent mastrevirus was identified. The complete mastrevirus genomes recovered here together with all those dicot-infecting mastrevirus publically available were run through SDT software in order to determine the pairwise identities. The eight CpCDV isolates recovered shared >84% genome-wide nucleotide sequence identity with previously described CpCDV isolates (Fig. 5.1). Based on guidelines proposed by Muhire *et al.* (2013) the eight isolates were identified as CpCDV strains D (n=5), C (n=2) and H (n=1) (Table 5.1). Interestingly, this is the first report of the D and H strains of CpCDV in Pakistan. Previously these strains have only been identified in India and Eritrea, respectively. Although CpCDV has previously been identified infecting a lentil plant in Pakistan (Kraberger *et al.*, 2013) this was a CpCDV-F isolate. CpCDV-C has so far only been found in Pakistan and, prior to the study here it was only isolated from chickpea.

Table 5.1: Sampling and isolate details of sequences recovered in this study.

Genbank accession #	Viral isolate	Host	Sampling location	Sampling year	Isolate ID
HG934858	CpCDV-C	Chickpea (<i>Cicer arietinum</i>)	Faisalabad, Pakistan	2010	NIAB-C
FR687960	CpCDV-D	Chickpea	Bahawalnagar, Pakistan	2008	BGR-3
KM377673	CpCDV-C	Lentil (<i>Lens culinaris</i>)	Faisalabad, Pakistan	2012	LE-E
KM377668	CpCDV-D	Chickpea	Faisalabad, Pakistan	2012	PK31
KM377670	CpCDV-D	Chickpea	Faisalabad, Pakistan	2012	PK37
KM377671	CpCDV-D	Lentil	Faisalabad, Pakistan	2012	PK43
KM377672	CpCDV-D	Chickpea	Faisalabad, Pakistan	2012	PK103
KM377669	CpCDV-H	Chickpea	Faisalabad, Pakistan	2012	PK32
KM377674	CpYDV	Chickpea	Faisalabad, Pakistan	2012	PK103
KM377675	CpYDV	Chickpea	Faisalabad, Pakistan	2012	PK37

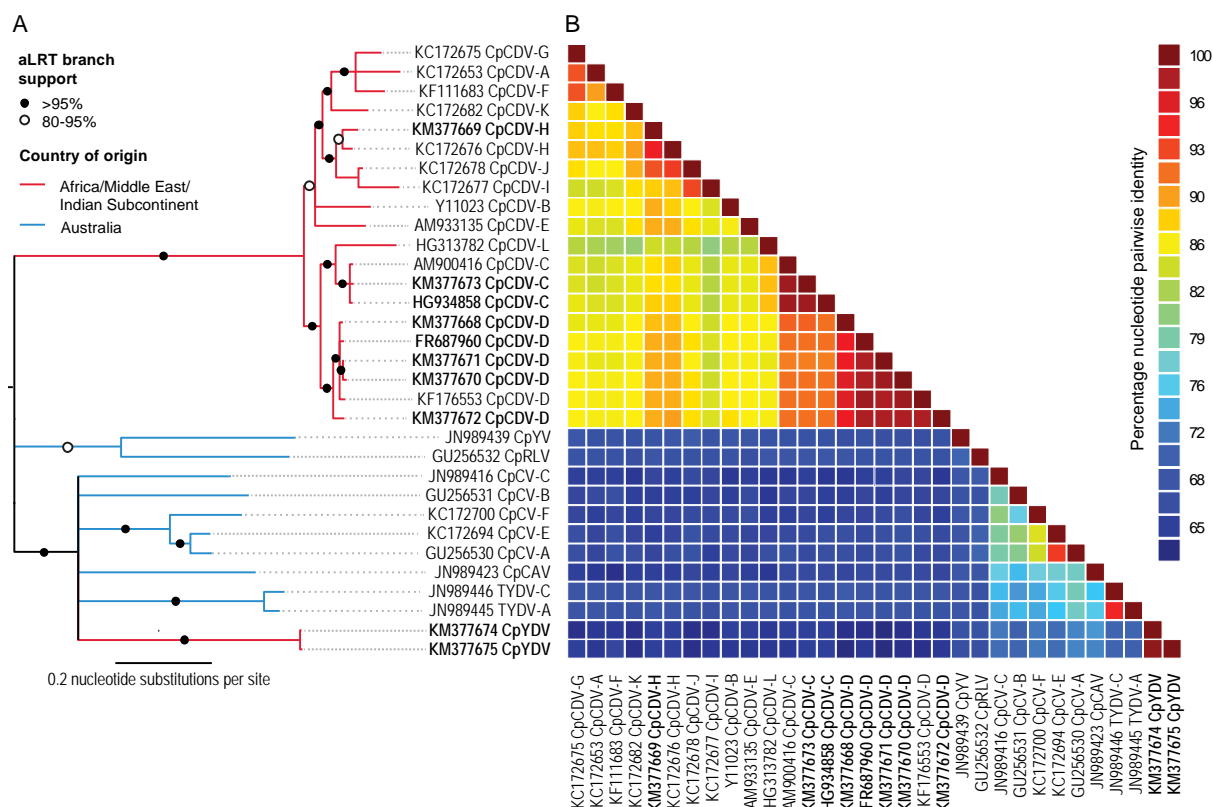


Figure 5.1: Maximum likelihood phylogenetic tree of dicot-infecting mastrevirus genomes determined in this study with representatives from each dicot-infecting mastrevirus strain and rooted with monocot-infecting mastreviruses. Mastrevirus isolates determined in this study highlighted in bold font. Branches are coloured by origin of samples and branches with <80% approximate likelihood branch support (aLRT) have been collapsed. B) Nucleotide pairwise identity colour matrix of dicot-infecting mastrevirus genomes. Mastrevirus isolates determined in this study are highlighted in bold font.

5.4.2 Discovery of two novel Australian-like mastrevirus isolates

In addition to the eight CpCDV isolates identified in this study, two divergent mastreviruses were recovered. The two isolates share 99.3% genome-wide nucleotide pairwise identity but <75% nucleotide pairwise identity to all other known mastreviruses (Fig. 5.1C). These isolates were obtained from two separate chickpea plants sampled in the vicinity of Faisalabad, central Punjab province. Both of these chickpea plants were also coinfecting with CpCDV-D (Table 5.1). Based on the mastrevirus species demarcation recommendations outlined by Muhire *et al.* (2013), coupled with phylogenetic support, these two viruses likely represent a new dicot-infecting mastrevirus species. This conclusion is supported by CP and Rep amino acid sequence phylogenetic analyses (Fig. 5.2A and B) which indicated that these novel viruses are more closely related to Australian dicot-infecting mastreviruses (with which they share ~69% CP and ~83% Rep pairwise amino acid identity) than they are to CpCDV (Fig. 5.2C). The name Chickpea yellow dwarf virus (CpYDV) is proposed for the new species. Full genome annotation indicates position of the MP, CP, Rep and RepA within the genome (Fig. 5.3). Further an amino acid annotation of each of these genes highlights the conserved domains and motifs found in geminivirus (Fig. 5.4). No evidence of recombination was found in the genome of CpYDV using the various recombination detection methods implemented in RDP4 (Martin *et al.*, 2010).

A global analysis of dicot-infecting mastreviruses has indicated that the centre of diversity and likely origin of the dicot-infecting mastreviruses is probably in or around Australia [12]. Our discovery of CpYDV in Pakistan suggests that the geographical origin of the dicot-infecting mastreviruses could be more difficult to pinpoint than previously thought. The present lack of information of CpYDV diversity means that it is not possible at this point to determine either where this species originated or whether it was recently introduced to Pakistan from elsewhere. There are several examples of plant viruses having been introduced over the past three decades into Pakistan, including: the nanovirus banana bunch top virus (Amin *et al.*, 2008), and the begomoviruses cotton leaf curl Gezira virus (Tahir *et al.*, 2011) and east Africa cassava mosaic virus (De Bruyn *et al.*, 2012). It is likely that these other viruses were introduced into the region around Pakistan within vegetatively propagated plant material and it remains a possibility that this could have been the route of entry for CpYDV – especially if the host range for this virus is as broad as that of CpCDV and includes crop or

ornamental plant species which are traded as vegetative material. The trade in ornamental plants, specifically *Hibiscus rosa-sinensis*, has been put forward as the most likely route for the introduction of a virus causing cotton leaf curl disease into China from the Indian subcontinent (Sattar *et al.*, 2013). However, it is similarly plausible that the geographic host range of either presently unsampled Australian CpYDV lineages or other Australian clade dicot-infecting mastreviruses is greater than is presently known, since no studies have so far looked for dicot-infecting mastreviruses in Southeast Asia.

It should also be pointed out, that the question of CpYDV's origins could be somewhat elucidated by the identification of its vector species. The CPs of mastreviruses are the sole determinant of insect vector specificity (Briddon *et al.*, 1990). Our analysis of the predicted CP amino acid sequences of the dicot-infecting mastreviruses shows that CpYDV groups within Australian dicot-infecting mastrevirus clade that contains TYDV (Fig. 5.2A). TYDV is the only virus in this clade for which a vector species has been identified; *Orosius orientalis* (syn. *O. argentatus*) (Trębicki *et al.*, 2010). CpCDV is reported to be transmitted by *O. orientalis* (Horn *et al.*, 1994) and *O. albicinctus* (Akhtar *et al.*, 2011) and the geographic host range of *O. albicinctus* extends from Australia across Southeast Asia to the Middle East (Wilson & Turner, 2010). Together this suggests that one or both of these *Orosius* spp. could also be the vector(s) of CpYDV: a possibility that would not help resolve the question of either CpYDV's origins or that of the dicot infecting mastreviruses as a whole. However, if it were found that CpYDV is transmitted by an Asian/Middle Eastern *Orosius* species that is not found in Australia, it would greatly support the hypothesis that this novel species originated outside Australia.

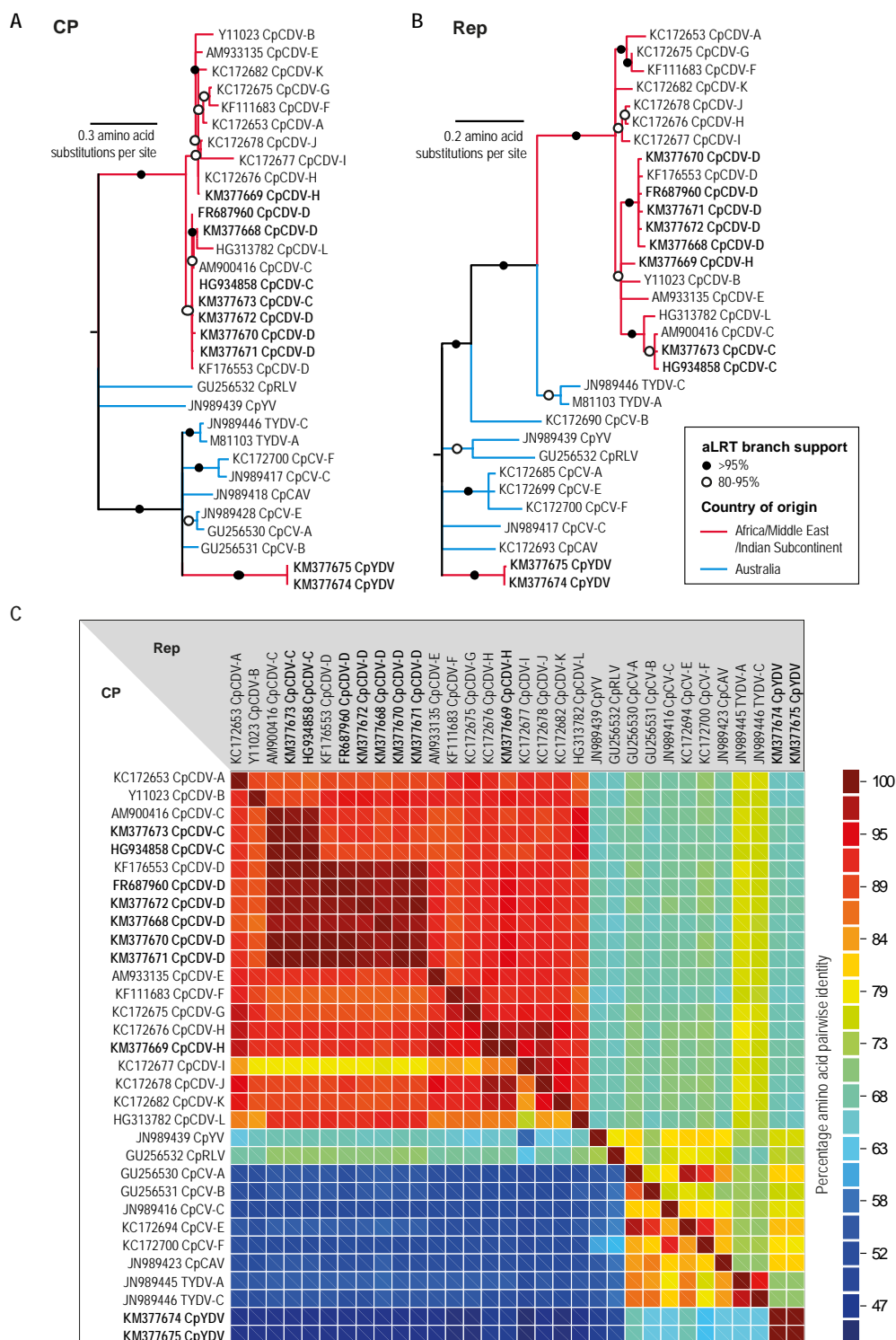


Figure 5.2: Maximum likelihood phylogenetic tree of the (A) CP and (B) Rep amino acid sequences of dicot-infecting mastreviruses. The trees were rooted with sequences of monocot-infecting mastreviruses. Mastrevirus isolates recovered in this study are highlighted in bold font. Branches are coloured origin of samples and branches with less than 80% approximate likelihood branch support (aLRT) have been collapsed. (C) Amino acid pairwise identity colour matrix of dicot-infecting mastrevirus Rep and CP sequences. Mastrevirus isolates determined in this study highlighted in bold font.

Chickpea yellow dwarf virus

Virion-strand origin of replication

TAATATTAC CAGTGAGTGCCTCTGCCCCACGCGAGCGCGTTAGCGCGAGCGGATTGGTTCTCGCAAAGTGAATCTCATGAGTGGATTATCTTGTCTATATAAAGGTGCTGCCTTTAATA [120]

Movement protein start codon

ATTTTCAAAGATGCTTCCCGCTAAATACCAAGTCTTTCTGGAGAAAATTACTCTTATACCCCACTTTTCGCGGGAAGTTATCAGGAAGTGCCTACCTCACATAATTCTTCTGGTGAGGC [240]

TTTTAACAAAGTCTTTGTGTCCTTATAGTTATTTTAATTTCCGTTGGTGTCTGTTATTTGGCTTACGTACTATTCTGTTAAGGATCTTATCTTCTATTGAAGGCTAAAAACAAGGAC [360]

Capsid protein start codon Movement protein stop codon

TACAACAGAGATAGGTTTTGGTAACACTCCAGGTAGACCTAACAGTCTGAAGACAAGAGGATGCGGGACCGTTTGGTAAAAACCTATTCAAGGAAGAAGGGTAAATATGCCAAGGCCT [480]

ACAAAGCTCTTGGTGTGAAAAATCAAAAAGAGCTTGAAGAGCTGGTAAATGCGCCTGCTTGTCCAGTTACTCTAGACCTGCCCTTACAGGTTGCTGAATATTTTTGGACTACGGACAAAA [600]

ATGGAATGATATTCGCTTCTGGTGGTGGTACTGCTCAATTTCACTATGTATCCTCAGGGTTCAAATGAGAATTGCAGACATTCCAACCAGACTAATACCTACAAGATGGCTATTAAGTGCT [720]

GGGTTGCATTGGATCCTACATTCTACAAGAAGGTGGCATGTGTTCTGTGCATTTCTGGTTGGTATACGACAAAGATCCGGGTAATACACTGCCTGGATGCTCTACCATTTTTGATACTC [840]

TGTATCAAGATTACCCGGGTACATGGACTGTATCCCGGAACGTTAGTCGTCGCTTTGTAGTTAAAAACATTTGGCACATAATCTTGGCCTCTAATGGAACGAATCCAACACAAGATCAAG [960]

ACCGTGTAAAGTATGCTGGACCTGGACCTGTGTTCAATGGAACACATGAACAAGTTTTTTAAAGACTTGGCGTAAGGACTGATTGGAAGAACTCAGCTACAGGTGAAGTAGCTGACA [1080]

Capsid protein stop codon

TTAAGAGTGGAGCATTGTACTTAGTTTGTGCACCAAGTGGTGGTGTCTGTGTAAGGGTTGGGGGTAGATTCAAGATGTACTTCAAATCCGTTGGAAATCAATAATTATTTTATTATTTT [1200]

GTACATATTCTTATCTTACAATATGACCTTGAAAAATAAATACATACAAAAAACACGCAAAAATGAACAAAAACAGAAAACAAACCTATATTTATGAAGTCTGAGTCAGAGGAGAGG [1320]

Replication-associated protein stop codon

CACGTTTAGTGACTCGACGCTGCCGAAGCAATAAAAGTCTCTCCTTCACTCATTATATGGATTTTACAATTTAAATAAAAGTATTCTCTCTGAGAATCTGACATACTCTCAAGCCAGTCC [1440]

RepA stop codon

TCATCATTTATTAATATTATAATACATGGGATTCTCTCTTAAACCTCTTCTTTTGGCGTATTTAGGGTAACTGTGAAATCCTTCTGTGACCCCTACTGCTTCCAGTTGGGGCAG [1560]

AACCTGAAGGGGATGTCGTGATGACGTTGTATGTTGCGTTGACGTCGTATGTAGTGAATCGACCCCTCCGTTGAAGTAGTTGTGTCTTCCAGACTTCTGGCCCAACTGGTCTTTCCG [1680]

Rep intron

GTACGACTTGGTCCGCAGATGTAGAGGGATCGTCGCCGAGTTGGAGTTGTTTCGGGTTCTGGGTATAATCGTCCATCCAAATCAAATCTTCAGTTGCAGTTTCTAAAGAAAGATGTGGAT [1800]

GGAGTAGTTGATAAGAATCTACACTTACAACATAGAGGTCTCTCTGGTATAGTTTAGCCATTCACTAATTTCTTCATGGCAGTGAAGAAATTCATCTGGAAATGGGCTTTGATAGGGTGG [1920]

CTGGAAGAGTTTTGCTTGCAGTGTATTCAAGCCAGTGAAGCTTAGTTGCCATTCACTGGGGGAATCTTGTGTTAATCATTTCCAGATACTCCTCTTTAGATGTTGCAGTCTGTATGATAG [2040]

TTCTCCATCGTTTCATCAGATTTGGTTGGGCTGACTCGATGGTTTCTGAAGTCACTCTGGTGTATGATGTTCCCGTCTTGGATATGTACTCAAGAACCCTGCTTTGAGTTCTTAGCAGGTT [2160]

GGATATTTGGATGGTTTTCTCATAAATCAAAAAAGAGGGTCCCGTATATCACATCTTTTGTCAAGCTGTATAAGACAGTGAAGATGGGGAGAACCCTCTTGGTGAAGTTCTAGTAGCAA [2280]

TAGCAATGAAAAAATAGCAAGAGATGTGAGTTTGGACCAGAGAAAACTCTGAGATTTCTGCAGTAGAGGAGCTATGTGGATAAGTAAGGAAACATATTTGTTTGAAGTCTGAAAG [2400]

Replication-associated and RepA protein start codon

TATA box

AGTTGTTGTTTGTACGTCTTGGCATGGTTCCTCAAAACAGTGTTCGAAAACCTCTATCTTAACCAAGTGAAGTGAAGGAAATGAGGAACAATATAGAGAGGTAATTGGGCCTGTTGGGCCT [2520]

GC-box – virion sense promoter element

GACACGTGGGCCGCACTCACTGAAGTT [2547]

Figure 5.3: Genome sequence annotation of CpYDV (GenBank accession numbers KM377674, KM377675).

Chickpea yellow dwarf virus

Movement protein

Predicted trans-membrane domain (Wright *et al.*, 1997)

MLPAKYQVFPGENYSYTPTFAGSYQEVPTSHNSSGETFNKVVFALIVILISVGVCLAYLVFKDLILLKAKKQRTTTEIGFGNTPGRPNRRQEDAGTV [101]

Capsid protein

Potential nuclear localisation signal

DNA binding domain

MPGPFSGKTYSRKKGKYAKAYKALGVKNQKELEELVNAPACPVTPRPALQVAEYFWTTDKNGMIFRSGGGTAHFTMYPQGSNENCRHSNQNTYKMAIKCWVALDPTFYKKVACVPVHFW [120]
LVYDKDPGNTLPGCSTIFDTLYQDYPGTWTVSRNVSRRFVVKHHWIIILASNGTNPTQDQDPAKYAGPGPVFQWKHMNKKFKRLGVRTDWKNSATGEVADIKSGALYLVCAPSGGAVVRV [240]
GGRFRMYFKSVGNQ [254]

Replication-associated protein

Rolling circle replication motifs I, II and III (4)

GRS domain (5)

MPRRNTNNSFRLQTKYVFLTYPHSSSTAENLRDFLWSKLTSLAIFFIATLHQDGSPLHLHCLIQLDKRCDIRDPSEFFDFEGNHPNIQPAKNSKQVLEYISKDGNIIITRGDFRNHRVSP [120]

Oligomerisation domain (3)

RBR interaction domain (6)

Walker-A (1, 2) (Gorbalenya & Koonin, 1993;

Gorbalenya *et al.*, 1990)

TKSDERWRTIIQTATSKEEYLEMIKQEFPHWATKLHWLEYASKLFPDIEPPYQSPFPDEFLLHCHEEITEWLNDRDLYVEPEQLQLRRRSLYICGPSRTGKTSWARSLSGRHNYFNNGGVDF [240]

Walker-B motif (1,2)

Motif C (1,2)

TTYDVNATYNVIDDI PFKFCPNWKQLVGSQKDFTVNPKYGKKRVKGGIPCTLIIVNNDWLESMSDSQREYFYLNCKIHIMSEGETFIASAASSH [336]

Figure 5.4: Annotated amino acid sequences of the MP, CP and Rep of CpYDV (GenBank accession numbers KM377674, KM377675).

1. Gorbalenya AE, Koonin EV, Wolf YI (1990) A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. FEBS letters 262:145-148
2. Gorbalenya AE, Koonin EV (1993) Helicases: amino acid sequence comparisons and structure-function relationships. Current Opinion in Structural Biology 3:419-429
3. Horváth G, Pettkó-Szandtner A, Nikovics K, Bilgin M, Boulton M, Davies J, Gutiérrez C, Dudits D (1998) Prediction of functional regions of the maize streak virus replication-associated proteins by protein-protein interaction analysis. Plant Molecular Biology 38:699-712
4. Koonin EV, Ilyina TV (1992) Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. The Journal of general virology 73 2763-2766
5. Nash TE, Dallas MB, Reyes MI, Buhrman GK, Ascencio-Ibañez JT, Hanley-Bowdoin L (2011) Functional analysis of a novel motif conserved across geminivirus Rep proteins. Journal of Virology 85:1182-1192
6. Xie O, Suarez-Lopez P, Gutierrez C (1995) Identification and analysis of a retinoblastoma binding motif in the replication protein of a plant

5.5 Concluding remarks

The results presented here extend our knowledge of the diversity and geographic distribution of dicot-infecting mastreviruses outside of Australia. We have revealed a greater degree of CpCDV diversity within Pakistan than was previously known and the existence outside Australia of the second dicot-infecting mastrevirus species. While this discovery slightly weakens the hypothesis that Australia is the geographical origin of dicot-infecting mastreviruses, it remains unclear whether CpYDV originated outside of Australia (perhaps in southeast Asia or the Middle East), or whether it was only recently introduced by humans to Pakistan from elsewhere. Certainly the occurrence of a divergent dicot-infecting mastrevirus outside of Australia reopens the question of the likely geographical origins of this interesting group of viruses.

GenBank accession numbers: KM377668 – KM377675

5.6 References

- Akhtar, K. P., Ahmad, M., Shah, T. M. & Atta, B. M. (2011).** Transmission of chickpea chlorotic dwarf virus in chickpea by the leafhopper *Orosius albicinctus* (Distant) in Pakistan -short communication. *Plant Protection Science* **47**, 1-4.
- Akhtar, S., Khan, A. J. & Briddon, R. W. (2013).** A Distinct Strain of Chickpea chlorotic dwarf virus Infecting Pepper in Oman. *Plant Disease* **98**, 286-286.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990).** Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.
- Amin, I., Qazi, J., Mansoor, S., Ilyas, M. & Briddon, R. W. (2008).** Molecular characterisation of Banana bunchy top virus (BBTV) from Pakistan. *Virus genes* **36**, 191-198.
- Briddon, R., Pinner, M., Stanley, J. & Markham, P. (1990).** Geminivirus coat protein gene replacement alters insect specificity. *Virology* **177**, 85-94.
- Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. (2011).** ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164-1165.
- De Bruyn, A., Villemot, J., Lefeuvre, P., Villar, E., Hoareau, M., Harimalala, M., Abdoul-Karime, A. L., Abdou-Chakour, C., Reynaud, B. & Harkins, G. W. (2012).** East African cassava mosaic-like viruses from Africa to Indian ocean islands: molecular diversity, evolutionary history and geographical dissemination of a bipartite begomovirus. *BMC evolutionary biology* **12**, 228.
- Edgar, R. C. (2004).** MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797.
- Farzadfar, S., Pourrahim, R., Golnaraghi, A. R. & Ahoonmanesh, A. (2008).** PCR detection and partial molecular characterization of Chickpea chlorotic dwarf virus in naturally infected sugar beet plants in Iran. *Journal of Plant Pathology* **90**, 247-251.
- Gorbalenya, A. E. & Koonin, E. V. (1993).** Helicases: amino acid sequence comparisons and structure-function relationships. *Current Opinion in Structural Biology* **3**, 419-429.
- Gorbalenya, A. E., Koonin, E. V. & Wolf, Y. I. (1990).** A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. *FEBS letters* **262**, 145-148.
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010).** New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**, 307-321.
- Hadfield, J., Thomas, J. E., Schwinghamer, M. W., Kraberger, S., Stainton, D., Dayaram, A., Parry, J. N., Pande, D., Martin, D. P. & Varsani, A. (2012).** Molecular characterisation of dicot-infecting mastreviruses from Australia. *Virus Research* **166**, 13-22.
- Hamed, A. A. & Makkouk, K. M. (2002).** Occurrence and management of Chickpea chlorotic dwarf virus in chickpea fields in northern Sudan. *Phytopathologia Mediterranea* **41**, 193-198.
- Horn, N. M., Reddy, S. V. & Reddy, D. V. R. (1994).** Virus-vector relationships of chickpea chlorotic dwarf geminivirus and the leafhopper *Orosius orientalis* (Hemiptera: Cicadellidae). *Annals of Applied Biology* **124**, 441-450.
- Kanakala, S., Verma, H. N., Vijay, P., Saxena, D. R. & Malathi, V. G. (2013).** Response of chickpea genotypes to Agrobacterium-mediated delivery of Chickpea chlorotic dwarf

- virus (CpCDV) genome and identification of resistance source. *Appl Microbiol Biotechnol* **97**, 9491-9501.
- Kraberger, S., Harkins, G. W., Kumari, S. G., Thomas, J. E., Schwinghamer, M. W., Sharman, M., Collings, D. A., Briddon, R. W., Martin, D. P. & Varsani, A. (2013).** Evidence that dicot-infecting mastreviruses are particularly prone to inter-species recombination and have likely been circulating in Australia for longer than in Africa and the Middle East. *Virology* **444**, 282-291.
- Liu, L., van Tonder, T., Pietersen, G., Davies, J. W. & Stanley, J. (1997).** Molecular characterization of a subgroup I geminivirus from a legume in South Africa. *Journal of General Virology* **78**, 2113-2117.
- Manzoor, M., Ilyas, M., Shafiq, M., Haider, M., Shahid, A. & Briddon, R. (2013).** A distinct strain of chickpea chlorotic dwarf virus (genus Mastrevirus, family Geminiviridae) identified in cotton plants affected by leaf curl disease. *Arch Virol* **159**, 1217-1221.
- Manzoor, M., Ilyas, M., Shafiq, M., Haider, M., Shahid, A. & Briddon, R. (2014).** A distinct strain of chickpea chlorotic dwarf virus (genus Mastrevirus, family Geminiviridae) identified in cotton plants affected by leaf curl disease. *Arch Virol* **159** (5), 1217-1221.
- Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D. & Lefevre, P. (2010).** RDP3: A flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**, 2462-2463.
- Morris, B. A. M., Richardson, K. A., Haley, A., Zhan, X. & Thomas, J. E. (1992).** The nucleotide sequence of the infectious cloned dna component of tobacco yellow dwarf virus reveals features of geminiviruses infecting monocotyledonous plants. *Virology* **187**, 633-642.
- Muhire, B., Martin, D. P., Brown, J. K., Navas-Castillo, J., Moriones, E., Zerbini, F. M., Rivera-Bustamante, R., Malathi, V. G., Briddon, R. W. & Varsani, A. (2013).** A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). *Arch Virol* **158**, 1411-1424.
- Mumtaz, H., Kumari, S. G., Mansoor, S., Martin, D. P. & Briddon, R. W. (2011).** Analysis of the sequence of a dicot-infecting mastrevirus (family Geminiviridae) originating from Syria. *Virus Genes* **42**, 422-428.
- Nahid, N., Amin, I., Mansoor, S., Rybicki, E., van der Walt, E. & Briddon, R. (2008).** Two dicot-infecting mastreviruses (family Geminiviridae) occur in Pakistan. *Arch Virol* **153**, 1441-1451.
- Posada, D. (2009).** Selection of models of DNA evolution with jModelTest. *Methods in molecular biology* **537**, 93-112.
- Sattar, M. N., Kvarnheden, A., Saeed, M. & Briddon, R. W. (2013).** Cotton leaf curl disease—an emerging threat to cotton production worldwide. *Journal of General Virology* **94**, 695-710.
- Tahir, M. N., Amin, I., Briddon, R. W. & Mansoor, S. (2011).** The merging of two dynasties – identification of an African cotton leaf curl disease-associated begomovirus with cotton in Pakistan. *PloS one* **6**, e20366.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. (2011).** MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**, 2713-2739.

- Thomas, J., Parry, J., Schwinghamer, M. & Dann, E. (2010).** Two novel mastreviruses from chickpea (*Cicer arietinum*) in Australia. *Arch Virol* **155**, 1777-1788.
- Trębicki, P., Harding, R., Rodoni, B., Baxter, G. & Powell, K. (2010).** Vectors and alternative hosts of Tobacco yellow dwarf virus in southeastern Australia. *Annals of Applied Biology* **157**, 13-24.
- Wilson, M. & Turner, J. (2010).** Handbook to leafhopper and planthopper vectors of plant disease. Edited by M. Wilson & J. Turner.
- Wright, E. A., Heckel, T., Groenendijk, J., Davies, J. W. & Boulton, M. I. (1997).** Splicing features in maize streak virus virion- and complementary-sense gene expression. *The Plant Journal* **12**, 1285-1297.

Chapter 6

Molecular diversity of *Chickpea chlorotic dwarf virus* in Sudan: high rates of intra-species recombination – a driving force in the emergence of new strains

Contents

6.1	Abstract.....	202
6.2	Introduction.....	203
6.3	Materials and methods.....	205
6.3.1	Sample collection, DNA isolation and full genome recovery	205
6.3.2	Sequence assembly and pairwise similarity calculations	206
6.3.3	Construction of phylogenetic trees	206
6.3.4	Recombination analysis.....	206
6.3.5	Selection analysis	207
6.4	Results and discussion	207
6.4.1	Strain classification of CpCDV isolates	207
6.4.2	The CpCDV population in Sudan is highly diverse.....	213
6.4.3	Evidence of extensive intra-strain recombination.....	218
6.4.4	Signals of natural selection within dicot-infecting mastrevirus species	225
6.5	Concluding remarks.....	232
6.6	References.....	235

This work has been published in *Infection, Genetics and Evolution*, and is presented in a similar manner to that of the publication:

Krabberger, S., Kumari, S., Hamed, A. A., Gronenborn, B., Thomas, J. E., Sharman, M., Harkins, G. W., Muhire, B. M., Martin, D. P., Varsani, A. (2014) Molecular diversity of *Chickpea chlorotic dwarf virus* in Sudan: high rates of intra-species recombination a driving force in the emergence of new strains. *Infection, Genetics and Evolution* 29, 203-215.

6.1 Abstract

In Sudan *Chickpea chlorotic dwarf virus* (CpCDV, genus *Mastrevirus*, family *Geminiviridae*) is an important pathogen of pulses that are grown both for local consumption, and for export. Although a few studies have characterised CpCDV genomes from countries within the Middle East, Africa and the Indian subcontinent, little is known about CpCDV diversity within any of the major chickpea production areas within these regions. Here we analyse the diversity of 145 CpCDV genomes sampled from pulses collected across the chickpea growing regions of Sudan. Although we find that seven of the twelve known CpCDV strains are present within the country, strain CpCDV-H alone accounted for ~73% of the infections analysed. Additionally we identified four new strains (CpCDV-M, -N, -O and -P) and show that recombination has played a significant role in the diversification of Sudanese CpCDV populations. Accounting for observed recombination events, we use the large amounts of data generated here to compare patterns of natural selection within protein coding regions of CpCDV and other dicot-infecting mastrevirus species.

6.2 Introduction

In Chapter Five a novel mastrevirus species was discovered together with eight isolates of CpCDV in a large sample set of symptomatic pulse material from Pakistan indicating a low incidence of CpCDV in the growing regions surveyed. The objective of the following study was to investigate further the dicot-infecting mastreviruses present in a major pulse growing region and gain an understanding of the dynamics on a more localised scale.

The Middle East, North Africa and the Indian subcontinent are all major producers of chickpeas, lentils, faba beans and various other pulses. In the Sudan pulses are both an important food source and a cash crop. They are grown in the fertile regions along the banks of the Nile which runs through the middle of the country, from South Sudan towards Egypt in the north. A serious constraint on pulse production in general, and on chickpea farming in particular, is the viral pathogen *Chickpea chlorotic dwarf virus* (CpCDV, genus *Mastrevirus*, family *Geminiviridae*). CpCDV is known to cause a variety of symptoms in chickpeas

including stunting, foliar yellowing or reddening and reduced seed production. Other important chickpea-infecting viruses are *Faba bean necrotic yellows virus* (genus *Nanovirus*, family *Nanoviridae*) and members of the genus *Polerovirus* (in the family *Luteoviridae*; (Abraham *et al.*, 2006; Kumari *et al.*, 2008; Makkouk *et al.*, 2003).

Globally there are seven known species of dicotyledonous plant-infecting mastreviruses (referred to here as dicot-infecting), five of which have only been documented in Australia: *Chickpea chlorosis virus*; CpCV (Hadfield *et al.*, 2012; Krabberger *et al.*, 2013; Thomas *et al.*, 2010), *Chickpea chlorosis Australia virus*; CpCAV (Hadfield *et al.*, 2012), *Chickpea redleaf virus*; CpRLV (Thomas *et al.*, 2010), *Chickpea yellows virus*; CpYV (Hadfield *et al.*, 2012) and *Tobacco yellow dwarf virus*; TYDV (Hadfield *et al.*, 2012; Morris *et al.*, 1992). The two species found outside of Australia are CpCDV (Ali *et al.*, 2004; Horn *et al.*, 1993; Krabberger *et al.*, 2013; Kumari *et al.*, 2004; Makkouk *et al.*, 1995; Manzoor *et al.*, 2014; Mumtaz *et al.*, 2011; Nahid *et al.*, 2008) and *Chickpea yellow dwarf virus* (CpYDV) (Krabberger *et al.*, 2014). Whereas CpCDV has been found in the Middle East (including Turkey), Africa and the Indian subcontinent, CpYDV has so-far only been found in Pakistan. With the exception of

TYDV, all of these dicot-infecting mastrevirus species have predominately been found infecting chickpeas, although little is known about their potential host range.

CpCDV is transmitted by the leafhopper species *Orosius orientalis* (Matsumura) (Horn *et al.*, 1993) and *O. albicinctus* (Distant) (Cicadellidae: Hemiptera) (Akhtar *et al.*, 2011; Kumari *et al.*, 2004). Natural hosts identified in the field include chickpea (*Cicer arietinum*) (Kraberger *et al.*, 2013; Kumari *et al.*, 2004; Mumtaz *et al.*, 2011; Nahid *et al.*, 2008), lentil (*Lens culinaris* Medik) (Kraberger *et al.*, 2013; Makkouk *et al.*, 2002), faba bean (*Vicia faba*) (Kraberger *et al.*, 2013), field pea (*Pisum sativum*) (Kraberger *et al.*, 2013), french bean (*Phaseolus vulgaris* L) (Ali *et al.*, 2004; Liu *et al.*, 1997), sugar beet (*Beta vulgaris* L) (Farzadfar *et al.*, 2008), *Sesbania bispinosa* (Jacq.) (Nahid *et al.*, 2008), pepper (*Capsicum annum* L.) (Akhtar *et al.*, 2013) and cotton (*Gossypium* sp.) (Manzoor *et al.*, 2014).

Recent studies have extended our current knowledge on CpCDV diversity (Kraberger *et al.*, 2013; Manzoor *et al.*, 2014; Mumtaz *et al.*, 2011) and there are currently twelve identified strains of CpCDV (A–L). Despite this there is little information on the prevalence and diversity of CpCDV within the pulse growing regions of individual countries. Field surveys in Sudan between 1996 and 2000 used serological analysis (tissue blot immunoassays) to reveal a CpCDV incidence of 72% in chickpea crops, and therefore identified this virus as the most important potential threat to chickpea production in Sudan (Hamed & Makkouk, 2002). The antibodies used for serological testing of CpCDV are also used to detect other dicot-infecting mastrevirus species and at the time no nucleotide sequence data was obtained for further analysis of these samples so therefore we cannot be confident that all samples detected as positive were in fact CpCDV.

Along with 16 CpCDV genomes from pulse samples collected between 1997 and 2008, we obtained full-length CpCDV genomes from an additional 129 samples collected in Sudan between 2012 and 2014 and two available in GenBank in order to analyse the molecular diversity for the first time of a large representative CpCDV sample set at high resolution within the pulse-growing areas of an entire country. Besides discovering four new CpCDV strains, we found evidence that extensive inter- and intra-strain recombination has made a major contribution to the diversification of this species. Finally, we capitalised on the large

amounts of data generated here to compare patterns of natural selection found in CpCDV to those found in other known monocot-infecting mastreviruses.

6.3 Materials and methods

6.3.1 Sample collection, DNA isolation and full genome recovery

In the growing seasons 2012-2014 leaf material from pulse plants displaying foliar yellowing, mosaic/mottling patterns and/or stunting was collected from 312 individual plants located in the major growing pulse-growing regions of Sudan. Additionally, pulse plant samples collected in Sudan in 1997, 2006 and 2008 were obtained from Institut des Sciences du Végétal, CNRS in France. A total of 312 samples from Sudan were screened from Gezira state (n=166), the River Nile state (n=141) and the Northern state (n=5). Also, opportunistic sampling of similarly symptomatic plants was undertaken in Morocco (n=18) in 2013.

Total genomic DNA from dried plant material was extracted using the GF-1 nucleic acid extraction kit (Vivantis Technologies, Malaysia), according to manufacturer's specifications. Circular DNA was enriched from DNA extracts using rolling circle amplification with the Illustra TempliPhi Amplification Kit (GE Healthcare, USA) as previously described (Owor *et al.*, 2007; Shepherd *et al.*, 2008). Full viral genomes were amplified from 0.5µl of enriched viral DNA using polymerase chain reaction (PCR). The PCR reaction comprised Kapa HiFi HotStart DNA polymerase (Kapa biosystems, USA) together with previously described degenerate back-to-back primers (dicot forward 5'-GAN TTG GTC CGC AGT GTA GA-3', dicot reverse 5'-GTA CCG GWA AGA CMW CYT GG-3') (Hadfield *et al.*, 2012). The thermal cycling protocol used was as follows: 94°C for 3 min, 25 cycles [98°C (3 min), 52°C (30 sec), 72°C (2:45 min)], 72°C for 3 min. PCR products were purified using the Quick-spin PCR Product Purification Kit (iNtRON Biotechnology, Korea) and ligated into pJET1.2 vector using the CloneJET™ PCR cloning kit (Fermentas, USA). Resulting recombinant plasmids containing viral genomes were sequenced at Macrogen Inc. (Korea) using primer walking.

6.3.2 Sequence assembly and pairwise similarity calculations

Complete CpCDV genomes were assembled using DNA Baser Sequence Assembler V4 (Heracle BioSoft, Romania) and probable genes identified using DNAMAN V7 (Lynnon Biosoft, Canada). The complete CpCDV genome sequences determined in this study together with 115 publically available dicot-infecting mastrevirus sequences, and a single wheat dwarf virus (WDV) genome (accession number AM040732) used to root the phylogenetic tree, (downloaded on 01 July 2014) were aligned using MUSCLE (Edgar, 2004). The resulting alignment was manually edited with MEGA5.2 (Tamura *et al.*, 2011). Percentage pairwise sequence similarities between CpCDV genome sequences were determined using SDT v1.0, calculated as 1–p-distance, with pairwise deletion of gaps (Muhire *et al.*, 2014).

6.3.3 Construction of phylogenetic trees

A full genome maximum likelihood phylogenetic tree was constructed using PHYML version 3 (Guindon *et al.*, 2010) using the best fit substitution model (TN93+G+I510) identified using jModelTest (Darriba *et al.*, 2012) and rooted with WDV. Branches with approximate likelihood ratio test (aLRT) support <80% were collapsed using Mesquite version 2.75 (Maddison & Maddison, 2011).

6.3.4 Recombination analysis

Two datasets, one containing all available dicot-infecting mastrevirus sequences and the other containing only the available CpCDV sequences were compiled and analysed for evidence of recombination using RDP4 (Martin *et al.*, 2010) with the following methods: RDP (Martin & Rybicki, 2000), GENECONV (Padidam *et al.*, 1999), Bootscan (Martin *et al.*, 2005), Maxchi (Smith, 1992), Chimera (Posada & Crandall, 2001), Siscan (Gibbs *et al.*, 2000), and 3Seq (Boni *et al.*, 2007). Recombination events were only accepted as credible when (1) they were detectable by three or more of these methods with associated p-values <10⁻³ and (2) they had strong phylogenetic support (i.e. clustering of the identified recombinant(s) in different parts of phylogenetic trees constructed from regions of the alignment corresponding to the fragments of the recombinant identified as having been derived from each of its parents).

6.3.5 Selection analysis

The full genome dicot-infecting mastrevirus dataset was divided into movement protein (MP), capsid protein (CP) and replication-associated protein (Rep) coding regions and realigned using codon information with MUSCLE. From these alignments we extracted separate CpCDV, CpCV, CpCAV and TYDV, for each coding region (MP, CP, and Rep) dataset. Accounting for recombination breakpoints identified with the GARD method (Kosakovsky Pond *et al.*, 2006) these 12 datasets were separately analysed for evidence of selection acting on individual codon sites using the MEME (Murrell *et al.*, 2012) and FUBAR (Murrell *et al.*, 2013) methods implemented in the HyPhy package via the online Datamonkey server (<http://www.datamonkey.org/>) (Delpont *et al.*, 2010). The FUBAR method was used to identify individual codon sites evolving under either diversifying or negative selection throughout the entire evolutionary history of the sequences being analysed and the MEME method was used to identify individual codons evolving under episodic diversifying selection within individual sub-lineages within the analysed datasets. There were too few sequences available for CpYDV, CpRV, CpRLV and CpYV for us to perform selection analyses on these species.

6.4 Results and discussion

6.4.1 Strain classification of CpCDV isolates

In this study a total of 145 full CpCDV genomes were recovered and sequenced from infected pulse plants collected in the Gezira state (n=104), the River Nile state (n=38) and the Northern state (n=3) of Sudan and from one sample obtained from Morocco (Table 6.1). Prior to this study twelve CpCDV strains had been identified from Syria, Turkey, Iran, South Africa, Pakistan, India, Sudan, Yemen and Eritrea. To our knowledge this is the first report of CpCDV in Morocco.

Based on the mastrevirus classification guidelines outlined by Muhire *et al.* (2013) we were able to assign 140 of the CpCDV isolates from this study to seven of the 12 previously described CpCDV strains through full genome analysis; C (n=18), D (n=3), E (n=1), F (n=3), H (n=107), I (n=1) and K (n=7). The remaining six isolates most likely represent four new

strains which we have tentatively named CpCDV-M (n=2), -N (n=1), -O (n=1) and -P (n=2). For the purpose of this study, we further classified CpCDV strains into genotype/variant groupings based on a threshold of 94 - 97% pairwise sequence identity (i.e. isolates with >97% pairwise sequence identity represent the same genotype, whereas those with >94% but <97% pairwise sequence identity represent different genotypes). The monophyly of these genotypes were also phylogenetically supported (Fig. 6.1 and Additional Table 6.1). Additionally, we further classified these isolates into variant groupings (i.e. groups of isolates with pairwise sequence identities of $\geq 99\%$ were classified as belonging to the same group of variants). Different variant groupings were denoted with a numerical subscript after the strain letter designator: e.g. CpCDV-H variant group one and two were denoted CpCDV-H₁ and CpCDV-H₂.

We noted that strains CpCDV-J and -I, include genomes that share more than 97% sequence identity and we therefore merged into strain CpCDV-I. Similarly, two CpCDV-G isolates were found to share >95% sequence identity with the CpCDV-F isolates and we therefore merged these isolates into the strain CpCDV-F.

Table 6.1: Sampling information for all CpCDV isolate sequences recovered in this study.

Strain	Host	Year sampled	Country	Region collected	Genbank no.
CpCDV-C	<i>Cicer arietinum</i>	2008	Sudan	El Talha	KM229768
	<i>Cicer arietinum</i>	2012	Sudan	Gezira	KM229780
	<i>Cicer arietinum</i>	2013	Sudan	Berber	KM229772
	<i>Cicer arietinum</i>	2013	Sudan	Berber	KM229773
	<i>Cicer arietinum</i>	2013	Sudan	Berber	KM229774
	<i>Cicer arietinum</i>	2013	Sudan	Berber	KM229775
	<i>Vicia faba</i>	2013	Sudan	Berber	KM229785
	<i>Cicer arietinum</i>	2013	Sudan	Berber	KM229771
	<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229769
	<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229770
	<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229777
	<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229781
	<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229782
	<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229783
	<i>Cicer arietinum</i>	2013	Sudan	Middle Gezira	KM229784
	<i>Cicer arietinum</i>	2014	Sudan	Berber	KM229776
	<i>Cicer arietinum</i>	2014	Sudan	Berber	KM229778
	<i>Cicer arietinum</i>	2014	Sudan	Berber	KM229779
CpCDV-D	<i>Pisum sativum</i>	2008	Sudan	ARC, Wad Medani	KM229786
	<i>Cicer arietinum</i>	2008	Sudan	Unknown	KM229787
	<i>Cicer arietinum</i>	2013	Morocco	Unknown	KM229788
CpCDV-E	<i>Vicia faba</i>	1997	Sudan	El Rayafa	KM229789
CpCDV-F	<i>Cicer arietinum</i>	2013	Sudan	Ad-Damar	KM229790
	<i>Cicer arietinum</i>	2013	Sudan	Ad-Damar	KM229791
	<i>Cicer arietinum</i>	2013	Sudan	Ad-Damar	KM229792
CpCDV-H	<i>Cicer arietinum</i>	2006	Sudan	Wad Medani	KM229801
	<i>Cicer arietinum</i>	2008	Sudan	ARC, Wad Medani	KM229795
	<i>Cicer arietinum</i>	2008	Sudan	ARC, Wad Medani	KM229796
	<i>Cicer arietinum</i>	2008	Sudan	ARC, Wad Medani	KM229797
	<i>Cicer arietinum</i>	2008	Sudan	ARC, Wad Medani	KM229798
	<i>Cicer arietinum</i>	2008	Sudan	ARC, Wad Medani	KM229799
	<i>Cicer arietinum</i>	2008	Sudan	El Talha	KM229794
	<i>Cicer arietinum</i>	2008	Sudan	Kalli region, North of Shendi	KM229800
	<i>Cicer arietinum</i>	2008	Sudan	Wad Medani	KM229793
	<i>Cicer arietinum</i>	2013	Sudan	Abuselaim	KM229885
	<i>Cicer arietinum</i>	2013	Sudan	Abuselaim	KM229887
	<i>Cicer arietinum</i>	2013	Sudan	Abuselaim	KM229888
	<i>Cicer arietinum</i>	2013	Sudan	Ad-Damar	KM229871
	<i>Cicer arietinum</i>	2013	Sudan	Ad-Damar	KM229873
	<i>Cicer arietinum</i>	2013	Sudan	Ad-Damar	KM229869
	<i>Cicer arietinum</i>	2013	Sudan	Ad-Damar	KM229870
	<i>Cicer arietinum</i>	2013	Sudan	Baika	KM229890
	<i>Cicer arietinum</i>	2013	Sudan	Berber	KM229857
	<i>Cicer arietinum</i>	2013	Sudan	Berber	KM229858
	<i>Cicer arietinum</i>	2013	Sudan	Berber	KM229864
	<i>Cicer arietinum</i>	2013	Sudan	Berber	KM229865
	<i>Cicer arietinum</i>	2013	Sudan	Berber	KM229866
	<i>Cicer arietinum</i>	2013	Sudan	Berber	KM229867
	<i>Vicia faba</i>	2013	Sudan	Berber	KM229899
	<i>Cicer arietinum</i>	2013	Sudan	Berber	KM229853
	<i>Cicer arietinum</i>	2013	Sudan	Berber	KM229854

Table 6.1 continued

<i>Cicer arietinum</i>	2013	Sudan	Berber	KM229855
<i>Cicer arietinum</i>	2013	Sudan	Berber	KM229856
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229802
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229803
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229804
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229805
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229806
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229807
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229808
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229809
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229811
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229812
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229813
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229814
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229815
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229816
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229817
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229818
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229819
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229820
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229821
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229822
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229823
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229824
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229825
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229826
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229827
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229828
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229829
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229830
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229831
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229832
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229833
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229834
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229841
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229845
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229852
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229859
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229868
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229872
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229877
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229878
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229879
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229880
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229881
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229886
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229897
<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229898
<i>Cicer arietinum</i>	2013	Sudan	ARC, Wad Medani	KM229874
<i>Cicer arietinum</i>	2013	Sudan	ARC, Wad Medani	KM229875
<i>Cicer arietinum</i>	2013	Sudan	ARC, Wad Medani	KM229876
<i>Cicer arietinum</i>	2013	Sudan	ARC, Wad Medani	KM229860
<i>Cicer arietinum</i>	2013	Sudan	ARC, Wad Medani	KM229861
<i>Cicer arietinum</i>	2013	Sudan	ARC, Wad Medani	KM229862

Table 6.1 continued

	<i>Cicer arietinum</i>	2013	Sudan	ARC, Wad Medani	KM229863
	<i>Lens esculenta</i>	2013	Sudan	Hudiba station	KM229810
	<i>Cicer arietinum</i>	2013	Sudan	Komor	KM229850
	<i>Cicer arietinum</i>	2013	Sudan	Komor	KM229851
	<i>Cicer arietinum</i>	2013	Sudan	Mealeag	KM229891
	<i>Cicer arietinum</i>	2013	Sudan	Mealeag	KM229893
	<i>Cicer arietinum</i>	2013	Sudan	Mealeag	KM229894
	<i>Cicer arietinum</i>	2013	Sudan	Middle Gezira	KM229835
	<i>Cicer arietinum</i>	2013	Sudan	Middle Gezira	KM229836
	<i>Cicer arietinum</i>	2013	Sudan	Middle Gezira	KM229837
	<i>Cicer arietinum</i>	2013	Sudan	Middle Gezira	KM229838
	<i>Cicer arietinum</i>	2013	Sudan	Middle Gezira	KM229839
	<i>Cicer arietinum</i>	2013	Sudan	Middle Gezira	KM229840
	<i>Cicer arietinum</i>	2013	Sudan	Middle Gezira	KM229842
	<i>Cicer arietinum</i>	2013	Sudan	Middle Gezira	KM229843
	<i>Cicer arietinum</i>	2013	Sudan	Middle Gezira	KM229844
	<i>Cicer arietinum</i>	2013	Sudan	Middle Gezira	KM229846
	<i>Cicer arietinum</i>	2013	Sudan	Middle Gezira	KM229847
	<i>Cicer arietinum</i>	2013	Sudan	Middle Gezira	KM229848
	<i>Cicer arietinum</i>	2013	Sudan	Middle Gezira	KM229849
	<i>Cicer arietinum</i>	2013	Sudan	Selaim	KM229882
	<i>Cicer arietinum</i>	2013	Sudan	Selaim	KM229883
	<i>Cicer arietinum</i>	2013	Sudan	Selaim	KM229884
	<i>Cicer arietinum</i>	2013	Sudan	Wad asha	KM229892
	<i>Cicer arietinum</i>	2013	Sudan	Wad Asha	KM229895
	<i>Cicer arietinum</i>	2013	Sudan	Wad elmaak	KM229896
	<i>Cicer arietinum</i>	2014	Sudan	Berber	KM229889
CpCDV-I	<i>Cicer arietinum</i>	2013	Sudan	Middle Gezira	KM229900
CpCDV-K	<i>Cicer arietinum</i>	1997	Sudan	Abu Haraz	KM229901
	<i>Cicer arietinum</i>	2008	Sudan	El Talha	KM229902
	<i>Cicer arietinum</i>	2008	Sudan	El Talha	KM229903
	<i>Cicer arietinum</i>	2013	Sudan	Gezira	KM229904
	<i>Cicer arietinum</i>	2013	Sudan	Komor Galeen	KM229906
	<i>Cicer arietinum</i>	2013	Sudan	Neemalaha	KM229907
	<i>Cicer arietinum</i>	2013	Sudan	Shalawa Galeen	KM229905
CpCDV-M	<i>Cicer arietinum</i>	2013	Sudan	Ad-Damar	KM229909
	<i>Cicer arietinum</i>	2013	Sudan	Ad-Damar	KM229908
CpCDV-N	<i>Cicer arietinum</i>	2013	Sudan	ARC, Wad Medani	KM229910
CpCDV-O	<i>Cicer arietinum</i>	2014	Sudan	Berber	KM229911
CpCDV-P	<i>Cicer arietinum</i>	2013	Sudan	Abuselaim	KM229912
	<i>Cicer arietinum</i>	2014	Sudan	Berber	KM229913

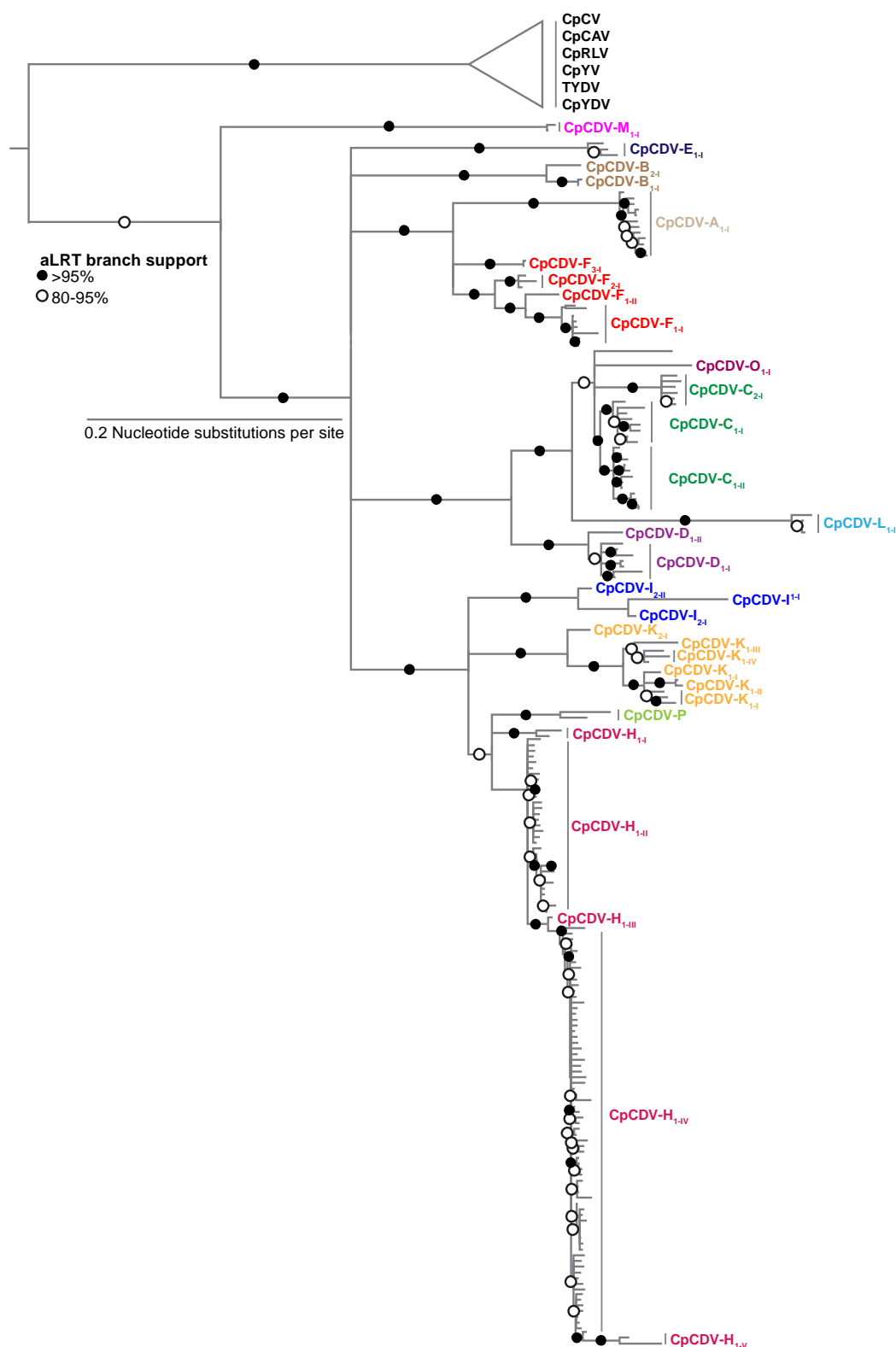


Figure 6.1: Maximum likelihood phylogenetic tree indicating the relationships of the known dicot-infecting mastrevirus species. Full genome sequences of CpCDV from this study together with those available in GenBank, as well as the five other species of dicot-infecting mastreviruses from Australia are included. Australian dicot-infecting mastrevirus species are represented by a triangle. Branch support is indicated by the open and closed circles in the key. Branches with less than 80% aLRT support have been collapsed.

6.4.2 The CpCDV population in Sudan is highly diverse

The CpCDV genomes determined here were primarily obtained from the major chickpea producing areas of Sudan. These regions were located along the Nile in the proximity of Wad Medani (Gezira State), Shendi and Berber (Nile State) and Selaim (Northern State) (Fig. 6.2). Prior to this study only two publically available full genomes of CpCDV (AM933134 and AM933135) had been determined (isolated from chickpeas) from Sudan. Both of these genomes belonged to strain E and both were sampled in 1997 from Abu Haraz, near Wad Medani (Gezira State). Here we have recovered an additional Sudanese CpCDV-E genome from a plant sample collected in 1997 El Rayafa (Fig. 6.3; Table 6.1). It is interesting to note that, despite the scale of our sampling, no other CpCDV-E isolates were recovered (Fig. 6.3 and Fig. 6.4). The absence of CpCDV-E variants in any samples collected here and elsewhere since 1997 suggests that this strain is a rare variant in the CpCDV population of Sudan.

Other CpCDV strains appear to be more persistent than CpCDV-E. CpCDV-K has so far only ever been found in the region surrounding Wad Medani, it was sampled there in 1997, 2008 and 2013. Similarly CpCDV-C and CpCDV-H isolates were consistently sampled in Sudan between 2006 and 2014. CpCDV-C has only ever been found in the Berber region, whereas CpCDV-H was found at every sampling site and accounted for 73% of all CpCDV isolates collected in Sudan since 2006. Given the prevalence of CpCDV-H in Sudan, it is unsurprising that this strain was also detected in neighbouring Eritrea in 2005 (Kraberger *et al.*, 2013).

Due to the low numbers of CpCDV genomic sequences that have been sampled outside of Sudan it is not currently possible to accurately infer the world-wide diversity and prevalence of the various CpCDV strains. Nonetheless, based on the available data, CpCDV-F is apparently the most widely distributed CpCDV strain in that it has been detected in six of the eleven countries where CpCDV genomes have been found.

Other strains appear to be confined to specific regions. For example, CpCDV-A has only been identified in the region encompassing Iran, Syria and Turkey. It is however, important to reiterate that there are only five countries (Sudan, India, Pakistan, Eritrea and Iran) from

which nine or more CpCDV genomes have been sampled and it is possible that extra sampling will reveal that the less commonly detected strains such as -B and -K have a greater geographical range than is presently apparent from the available data.

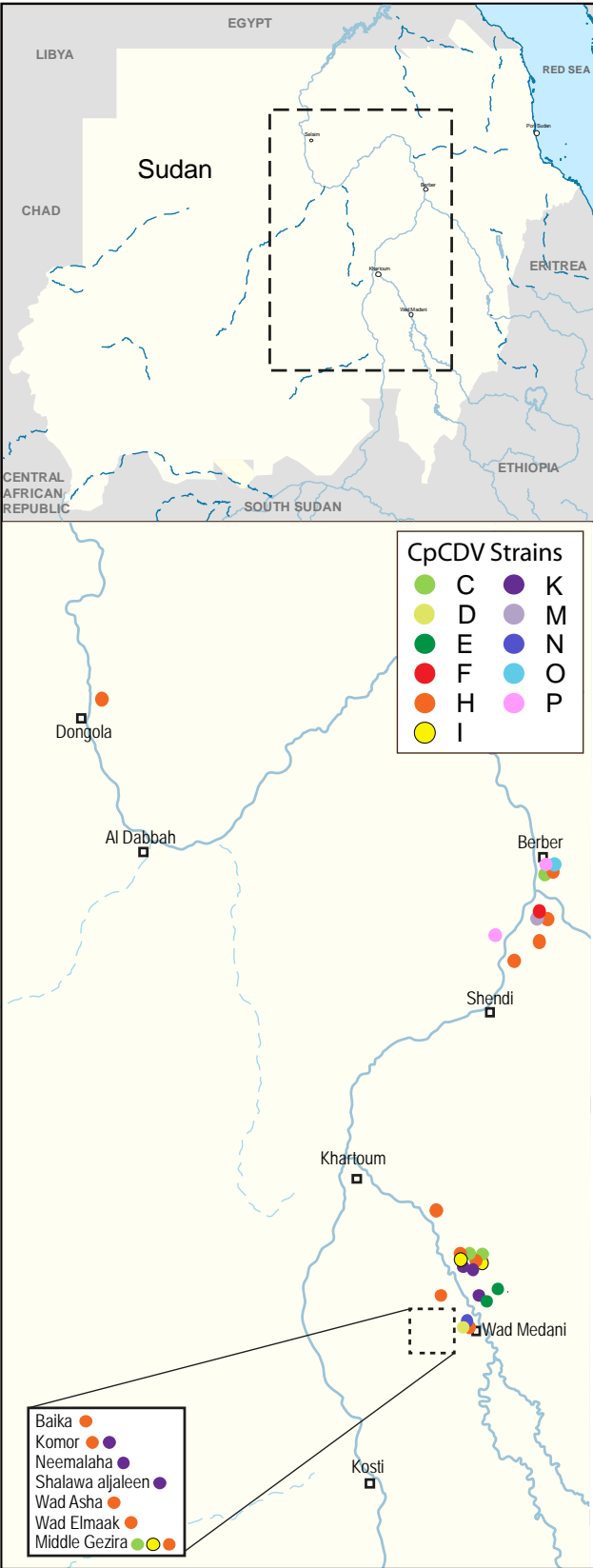


Figure 6.2: Sampling sites and the distribution of CpCDV strains in Sudan.

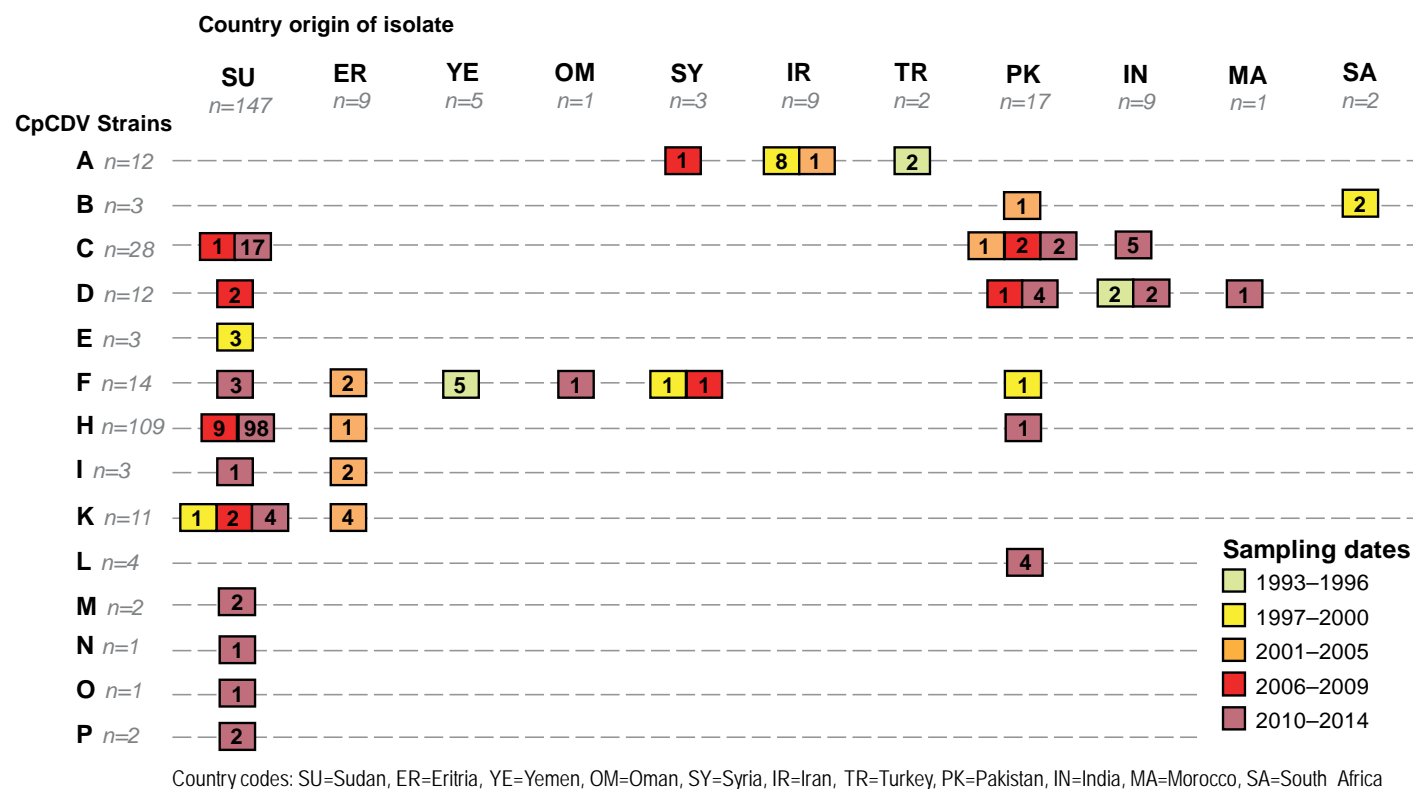


Figure 6.3: Summary of country of origin, strain, sampling year and total numbers of CpCDV genomes identified including those recovered in this study and all other available sequences in GenBank. Sampling dates have been clustered into four year intervals which are indicated by the gradient of colours shown in the key. Number of genome sequences determined for each strain and country are indicated within each corresponding box.

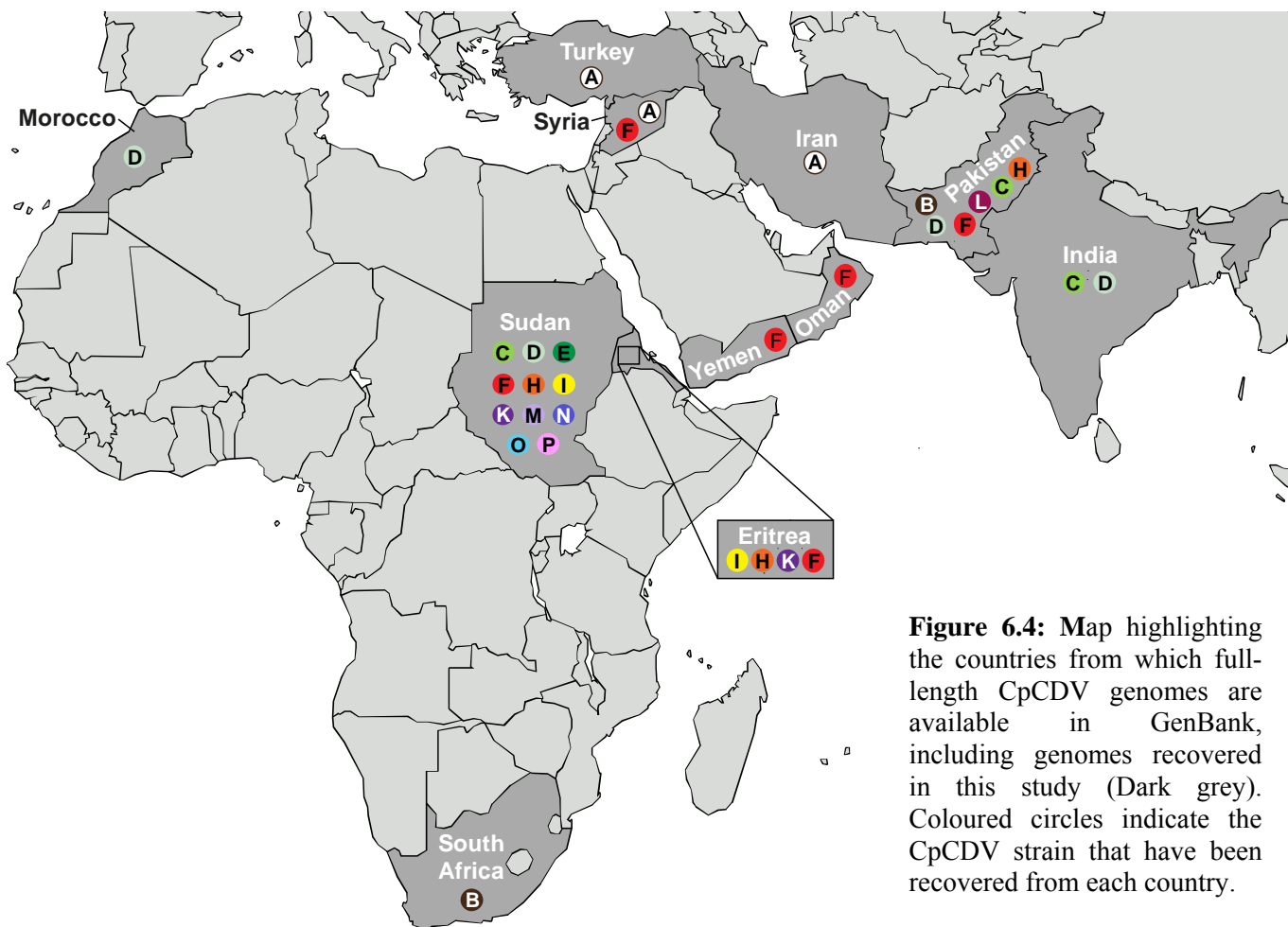


Figure 6.4: Map highlighting the countries from which full-length CpCDV genomes are available in GenBank, including genomes recovered in this study (Dark grey). Coloured circles indicate the CpCDV strain that have been recovered from each country.

6.4.3 Evidence of extensive intra-strain recombination

As with other geminiviruses, recombination is apparently a major feature of mastrevirus evolution. We analysed the 146 genomes from this study together with other dicot-infecting mastrevirus genomes available in GenBank for evidence of recombination. Previously many inter-strain recombination events (Kraberger *et al.*, 2013) and inter-species recombination events (Hadfield *et al.*, 2012; Kraberger *et al.*, 2013; Martin *et al.*, 2011b) involving CpCDV have been identified. We therefore attempted to both characterise novel recombination events evident within the genome sequences determined here and refine the characterisations of previously identified recombination events involving CpCDV.

We detected evidence of two unique intra-strain, 29 unique inter-strain (intra-strain and inter-strain events are collectively referred to as intra-species events) and six unique inter-species recombination events, all of which involved CpCDV isolates as sequence acceptors (Fig. 6.5). Of these events 19 intra-species events and one inter-species recombination event have not previously been identified. For four of the ten previously detected intra-species recombination events (Hadfield *et al.*, 2012; Kraberger *et al.*, 2013; Martin *et al.*, 2011b) we were able to for the first time, identify both of the likely parental sequences to at least the strain level (events 8, 12, 18 and 21 in Fig. 6.5A; Table 6.2).

Recombination appears to have played a particularly predominant role in the genesis of strains CpCDV-N, -O and -P (three of the novel strains identified here for the first time). The ancestral progenitor of each of these strains was likely derived from several recombination events involving parental sequences belonging to other CpCDV strains found in Sudan. One event inferred to have occurred in the common ancestor of CpCDV-N and -O involves a parental sequence that is most similar to CpCDV-L (Event 18 in Fig. 6.5A; Table 6.2). Although CpCDV-L has so far only been found infecting cotton in Pakistan, it is entirely plausible that the CpCDV-L-like parent of the ancestral recombinant that yielded the O and N strains could have existed almost anywhere within (or even outside of) the presently known geographical range of CpCDV.

Of the 146 sequences classified here as belonging to CpCDV-H (the predominant CpCDV strain found in Sudan), 107 are recombinants with parental viruses likely belonging to CpCDV strains -I, -O, -C, -D and a currently unsampled strain (Events 9, 10, 11, 13, 27 and 29 in Fig. 6.5A; Table 6.2). Interestingly all but one of these recombination events (the exception being event 27) involved a CpCDV-H acquiring a *rep* gene fragment from one of the other CpCDV strains that are found in Sudan.

In fact, many of the other detected inter-strain recombination events have also involved the transfer of *rep* gene fragments. This pattern of sequence exchange mirrors that seen in other geminiviruses (specifically those in the *Curtovirus*, *Mastrevirus* and *Begomovirus* genera). Specifically, the recombination patterns evident here and elsewhere indicate that for geminiviruses in general, recombination frequencies may be higher in genome regions encoding complementary sense genes than they are in the regions encoding virion sense genes. Alternatively, if basal recombination frequencies are similar across the genome, it would imply that selection might generally disfavour the survival of recombinants with breakpoints within the virion sense genes more than it disfavours recombinants with breakpoints in the complementary sense genes (Kraberger *et al.*, 2012; Lefeuvre *et al.*, 2009; Martin *et al.*, 2011a; Martin *et al.*, 2011b; Varsani *et al.*, 2008).

Four of the six inter-species events that we detected involved the transfer of large genome fragments ranging in size from 753 to 1314 nt. (Events A, B, C and D in Fig. 6.5B; Table 6.3). Overall the inter-species recombination events involve the transfer of on average 21% of the genome. This percentage is substantially larger than those inferred in previous analyses of both dicot- and monocot-infecting mastreviruses (Hadfield *et al.*, 2012; Kraberger *et al.*, 2013; Kraberger *et al.*, 2012; Martin *et al.*, 2011b; Monjane *et al.*, 2011; Varsani *et al.*, 2009; Varsani *et al.*, 2008).

CpCDV, which is one of only two dicot-infecting mastrevirus species known to occur outside of Australia, was identified here as potentially being a parent (or at least being most closely related to the actual parent) in four of the six inter-species events; all of which involved sequence transfers to viruses currently found in Australia. This information, together with the discovery of the Australian-like mastrevirus, CpYDV, in Pakistan, implies that an ancestor of

CpCDV may have moved out of Australia (or at least that region of the world) and into the Middle East, Africa and the Indian subcontinent where it subsequently became established; a hypothesis supported by the phylogeographic analysis undertaken by Krabberger *et al.* (2013).

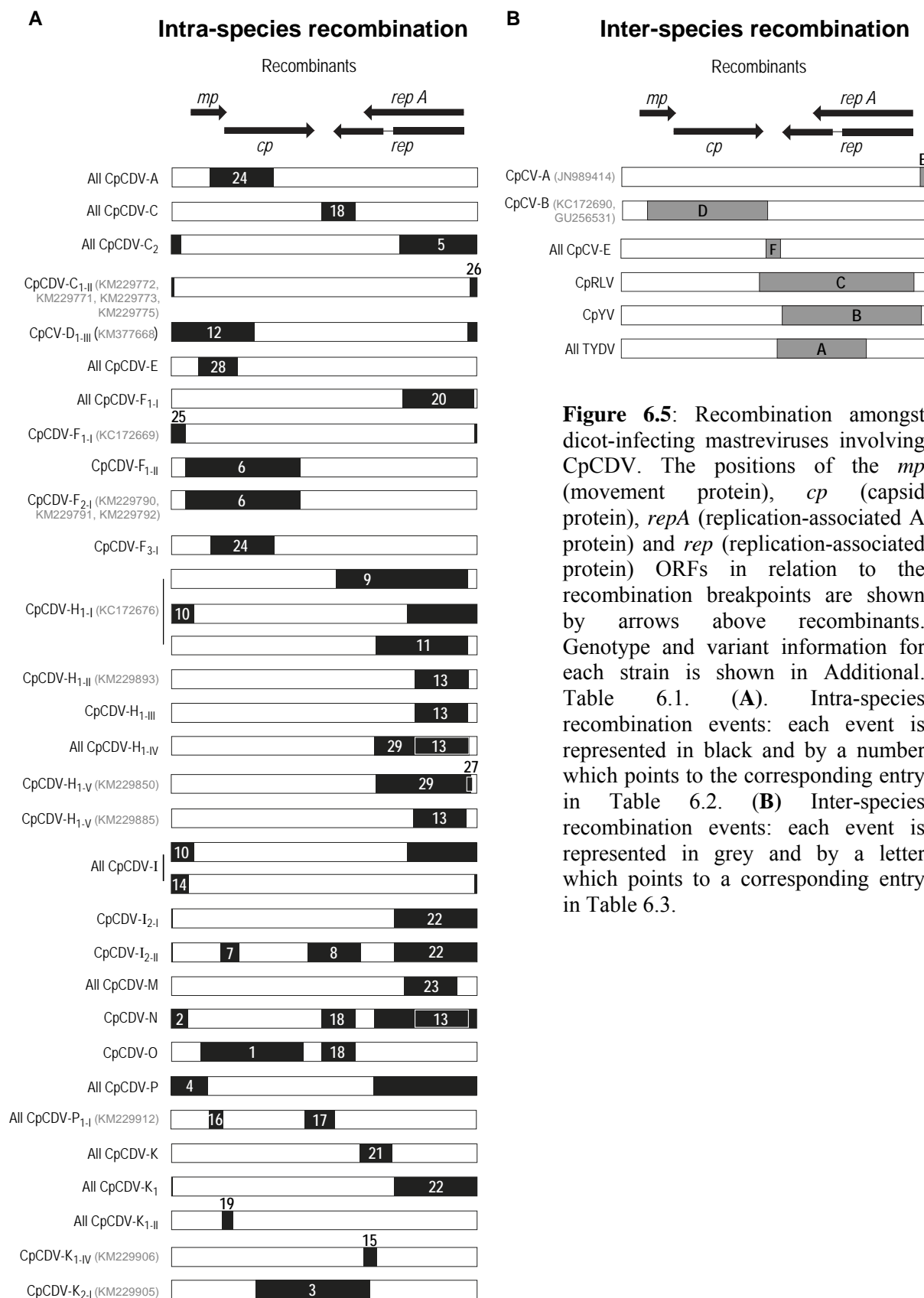


Figure 6.5: Recombination amongst dicot-infecting mastreviruses involving CpCDV. The positions of the *mp* (movement protein), *cp* (capsid protein), *repA* (replication-associated A protein) and *rep* (replication-associated protein) ORFs in relation to the recombination breakpoints are shown by arrows above recombinants. Genotype and variant information for each strain is shown in Additional Table 6.1. (A). Intra-species recombination events: each event is represented in black and by a number which points to the corresponding entry in Table 6.2. (B) Inter-species recombination events: each event is represented in grey and by a letter which points to a corresponding entry in Table 6.3.

Table 6.2: Summary of intra-species recombination events. Major and minor parent labels indicate the inferred parent(s) respectively donating the larger and smaller fraction of the recombinant's genome. The method with the most significant p-value is indicated in bold and the associated p-value is shown.

Event	Recombinant region	Potential major Parent	Potential minor Parent	Detection method	P-value
Intra-species recombination					
1	234-1096	All CpCDV-C, CpCDV-D _{1-II} , CpCDV-D _{1-III} , CpCDV-D _{1-I} (KM377671, KM229788, KC172664), CpCDV-N	All CpCDV-K, All CpCDV-P	RGBCST	9.42x10 ⁻⁴⁷
2	1689-138	All CpCDV-C, CpCDV-O	All CpCDV-H (except KM229850 and KM229885)	RGMCST	2.70x10 ⁻⁵⁷
3	708-1668	All CpCDV-K _{1-IV} , All CpCDV-K _{1-I} , CpCDV-K _{1-III}	All CpCDV-H (except KM229850), CpCDV-I _{2-II}	RGBMCST	5.60x10 ⁻⁶²
4	1700-310	All CpCDV-K _{1-I} , CpCDV-K _{1-III} , All CpCDV-K _{1-IV} , CpCDV-K _{2-I}	All CpCDV-H, All CpCDV-I ₂	RGBMCST	9.01x10 ⁻⁴⁴
5	1942-2543	All CpCDV-C ₁ , CpCDV-D _{1-I} (KM377674, KM377671, FR687960, KF176553, KC172664, KC172665), CpCDV-D _{1-II} , CpCDV-D _{1-III} , CpCDV-O	All CpCDV-K, CpCDV-I _{1-I}	RGBMCST	6.18x10 ⁻²⁶
6	119-1085	All CpCDV-F ₃	CpCDV-F _{1-I} (KC172667, KC172669, KC172670, KC172671)	RGBMCST	6.86x10 ⁻³⁸
7	404-564	CpCDV-I ₁ , CpCDV-I _{2-I}	All CpCDV-H	RGBMCST	7.88x10 ⁻¹⁵
8	1133-1583	CpCDV-I ₁ , CpCDV-I _{2-I}	All CpCDV-H, CpCDV-K ₂	RGBMCST	3.72x10 ⁻²⁷
9	1372-2485	All CpCDV-H _{1-III} , All CpCDV-H _{1-IV}	All CpCDV-I ₂	RGBMCST	1.93x10 ⁻¹⁵
10	1982-187	CpCDV-C _{1-I} (AM900416), CpCDV-C _{1-II} (KM229774, KM229776, KM229778, KM229779), CpCDV-D _{1-II} (KM229787), CpCDV-H _{1-II} (KM229893), CpCDV-H _{1-III} , CpCDV-H _{1-IV} (KM377669, KM229803, KM229806, KM229810, KM229819, KM229820, KM229824, KM229825, KM229827, KM229830, KM229836, KM229838, KM229840, KM229842, KM229844, KM229845, KM229847, KM229849, KM229851, KM229852, KM229860, KM229866, KM229869, KM229871, KM229872, KM229873, KM229878, KM229884, KM229888, KM229889, KM229890, KM229891, KM229895, KM229898, KM229793, KM229797)	All CpCDV-K	RGBMCST	1.47x10 ⁻¹²
11	1709-2483	CpCDV-H _{1-III} (KM229877), CpCDV-H _{1-III} , All CpCDV-H _{1-IV}	CpCDV-I _{2-I} , CpCDV-I _{2-II}	RGBMCST	4.49x10 ⁻¹³
12	2493-698	CpCDV-D _{1-I} (KC172664)	All CpCDV-C	RGBMCST	2.02x10 ⁻²¹
13	2029-2479	All CpCDV-P, All CpCDV-H _{1-II} (except KM229897)	Ancestral CpCDV-I-like	RGBMCST	9.47x10 ⁻¹²
14	2548-147	All CpCDV-H (except KM229850 and KM229885), CpCDV-P _{1-I} , CpCDV-P _{1-II}	All CpCDV-K	RGBMCST	1.73x10 ⁻⁰⁹
15	1602-1713	All CpCDV-K _{1-I} , CpCDV-K _{1-III} , CpCDV-K _{1-IV} (KM229904)	CpCDV-H _{1-I} (KM229843), All CpCDV-H _{1-II} (except KM229770), CpCDV-H _{1-III} , All CpCDV-H _{1-IV} (except KM229895), All CpCDV-H _{1-V}	GBLT	1.44x10 ⁻⁰⁹
16	305*-425*	CpCDV-K _{1-I} (KC172682)	CpCDV-H _{1-V} (KM229850), All CpCDV-H _{1-II} (except KM229815, KM229816, KM229817), All CpCDV-H _{1-IV} (except KM229889, KM229890, KM229884, KM229848, KM229838, KM229810, KM377669)	RBT	1.59x10 ⁻⁰³
17	1123-1371	All CpCDV-K	All CpCDV-E	RGBMCST	1.43x10 ⁻¹⁷
18	1258-1547	All CpCDV-D _{1-I} , CpCDV-D _{1-II}	All CpCDV-L	RGBMCST	4.14x10 ⁻⁰⁸
19	433-524	All CpCDV-K _{1-I} , CpCDV-K _{1-III} , All CpCDV-K _{1-IV} , CpCDV-K ₂ , CpCDV-O, All CpCDV-P	All CpCDV-E	RGBM	2.52x10 ⁻⁰⁸

Table 6.2 continued

20	1949-2548	Ancestral CpCDV-A-like	All CpCDV-F _{3-I}	RGBMCST	9.66x10 ⁻⁰⁷
21	1579-1850	All CpCDV-H _{1-II} (except KM229816, KM229876, KM229877), CpCDV-H _{1-III} , All CpCDV-H _{1-IV} (except KM229797, KM377669, KM229812, KM229813, KM229824, KM229834, KM229836, KM229837, KM229844, KM229845, KM229849, KM229851, KM229852, KM229861, KM229867, KM229872, KM229873, KM229878, KM229879)	All CpCDV-M	RBMC	1.05x10 ⁻⁰⁵
22	1873-6*	All CpCDV-E	All CpCDV-F ₃ , All CpCDV-A ₁ (except KC172654)	RGBMC S	6.62x10 ⁻⁰⁸
23	1957-2406	All CpCDV-F _{2-I} , All CpCDV-F ₃ , CpCDV-F _{1-II}	Ancestral CpCDV-P-like, CpCDV-H-like	GM S	9.07x10 ⁻¹⁰
24	330*-867	Ancestral CpCDV-I-like	All CpCDV-H _{1-III} , All CpCDV-H _{1-IV}	RBMC ST	1.01x10 ⁻⁰⁷
25	2549*-118*	CpCDV-F _{1-I} (KC172667)	Ancestral CpCDV-F _{2-I} -like	R BT	7.00x10 ⁻⁰⁵
26	2510-23	All CpCDV-C _{1-I} , CpCDV-C _{1-II} (except KM229771, KM229773, KM229772, KM229775), All CpCDV-D _{1-I} (except KM229786, KM377672), CpCDV-D _{1-II} , CpCDV-O	All CpCDV-M	R GB	1.41x10 ⁻⁰⁹
27	2514*-2562	CpCDV-H _{1-IV} (KM229886, KM229857, KM229850, KM229847, KM229811, KM229797, KM229800), All CpCDV-H _{1-II} (except KM229804, KM229807, KM229814, KM229815, KM229817, KM229822, KM229862, KM229864, KM229870, KM229877, KM229893, KM229894, KM229897, KM229859) CpCDV-H _{1-III} , CpCDV-H _{1-V} (KM229885)	CpCDV-D _{1-I} (KC172664, FR687960, KC172665)	R GBL	1.68x10 ⁻²⁸
28	206-536	All CpCDV-E	Ancestral CpCDV-C ₁ -like and CpCDV-D ₁ -like	M S	5.40x10 ⁻⁰⁴
29	1708*-2508*	All CpCDV-H _{1-II} , CpCDV-H _{1-III} , All CpCDV-H _{1-IV} (except KM229879), CpCDV-P _{1-I}	All CpCDV-C ₁ , All CpCDV-C _{2-I} (except KM229770) CpCDV-O	RBMC ST	1.35x10 ⁻⁴⁰

RDP (R) GENCONV (G), BOOTSCAN (B), MAXCHI (M), CHIMERA (C), SISCAN (S), LARD (L) and 3SEQ (T).

* = The actual breakpoint position is undetermined.

Table 6.3: Summary of inter-species recombination events. Major and minor parent labels indicate the inferred parent(s) respectively donating the larger and smaller fraction of the recombinant's genome. The method with the most significant p-value is indicated in bold and the associated p-value is shown.

Event	Recombinant region	Potential major Parent	Potential minor Parent	Detection method	P-value
Inter-species recombination					
A	1311-2064	All CpCAV	All CpCDV	RGBMCST	3.05×10^{-14}
B	1348-2517	All CpCDV	All CpCAV	RBMCS	6.82×10^{-11}
C	1168-2482	All CpCDV-B, All CpCDV-E, CpCDV-D ₁₋₁ (KM229787, KM377671, Q4790, KC172664, KC172665), All CpCDV-C ₁₋₁ , All CpCDV-F ₂₋₁ , All CpCDV-F ₁₋₁ , All CpCDV-F ₁₋₁ , All CpCDV-L, All CpCDV-M, All CpCDV-A (except KC172657, KC172658, KC172661), All CpCDV-H ₁₋₁ , CpCDV-H ₁₋₁ (KM229793, KM229794, KM229796, KM229797, KM229799, KM377669, KM229802, KM229803, KM229806, KM229809-11, KM229813, KM229819, KM229820, KM229825, KM229830, KM229836, KM229839-42, KM229844, KM229845, KM229847, KM229848, KM229851, KM229854, KM229856, KM229857, KM229860, KM229861, KM229866, KM229868, KM229869, KM229872, KM229878, KM229879, KM229880, KM229882, KM229884, KM229886, KM229889, KM229890, KM229891, KM229895, KM229896, KM229898)	All CpCAV (except KC172687 and KC172688)	RBMCS	6.62×10^{-10}
D	209-1234	Ancestral CpRV	All CpCV-E, CpCV-A (GU256530)	RBMCS	4.17×10^{-24}
E	2512-2556	CpCV-A (JN989415, GU256530, JN989413, KC172684), CpCV-E (JN989431, JN989433)	All CpCDV-B ₂₋₁ , All CpCDV-C ₁₋₁ , All CpCDV-C ₁₋₁ , All CpCDV-D (except KC172665 and KC172664), All CpCDV-H ₁₋₁ , All CpCDV-I ₁₋₁ , All CpCDV-K ₁₋₁ , All CpCDV-K ₁₋₁ , CpCDV-N, CpCDV-O, CpCDV-E (KM229901 and AM933135), CpCDV-B ₁₋₁ (AM849096), CpCDV-C ₂₋₁ (KM229800), CpCDV-H ₁₋₁ (KM229893)	RGBM	2.09×10^{-09}
F	1230-1335*	CpCV-F (KC172700)	Ancestral CpCDV-H-like	MST	4.60×10^{-06}

RDP (R) GENCONV (G), BOOTSCAN (B), MAXCHI (M), CHIMERA (C), SISCAN (S), LARD (L) and 3SEQ (T).

* = The actual breakpoint position is undetermined.

6.4.4 Signals of natural selection within dicot-infecting mastrevirus species

The large amount of novel CpCDV sequence data that we generated provided a good opportunity to compare and contrast signals of natural selection acting on the coding sequences of the various well sampled species of dicot-infecting mastreviruses. We used two different codon-model based methods, FUBAR and MEME, to infer selective processes acting on individual codon sites within the *mp* (Fig. 6.5A), *cp* (Fig. 6.5B) and the *rep* (Fig. 6.6) genes of all available CpCDV (n=205), CpCAV (n=13), CpCV (n=28) and TYDV (n=9) full-length genome sequences.

Due to the low numbers of available CpCAV and TYDV sequences these analyses only had sufficient power to detect statistically significant signals of natural selection at a few individual codon sites in each of these species.

Whereas dN/dS values significantly lower than one imply negative or purifying selection favouring the maintenance of amino acid sequences, dN/dS values significantly greater than one imply positive or diversifying selection favouring the modification of amino acid sequences. It is expected that most expressed viral proteins should have close to functionally optimal amino acid sequences and that their genes should therefore be evolving under predominantly negative selection (Duffy & Holmes, 2008; Krabberger *et al.*, 2012; Shackelton *et al.*, 2005; Stenzel *et al.*, 2014). It is unsurprising then that all of the analysed dicot-infecting mastrevirus genes had dN/dS values that were significantly lower than one, with the values for the *mp* generally displaying the lowest degree of purifying selection (i.e. the highest dN/dS) and the *cp* the highest degree (i.e. the lowest dN/dS). It is important to note here that whereas it is valid to compare the magnitudes of the dN/dS values between different genes of the same species, unless the datasets for the different species being analysed have similar degrees of diversity, it is not valid to compare the magnitudes of dN/dS values for the same gene in different species. In this regard, it is apparent that for all species the *cp* is evolving under stronger purifying selection than the *rep*. Also, with the exception of CpCAV, the *mp* is evolving under the weakest negative selection. Curiously, with CpCAV the *mp* is apparently evolving under the strongest negative selection. It should be noted, however, that the CpCAV dataset was both less diverse, and contained fewer sequences, than the other

datasets analysed: factors which both strongly impact the power of the analyses we have performed.

The influence of dataset size and diversity is also clearly reflected in the differences between the datasets with respect to the numbers of codon sites detectably evolving under either positive or negative selection. We were nevertheless able to detect a number of individual codon sites that appear to be consistently evolving under negative selection in two or more of the analysed species (sites indicated in red and orange in Fig. 6.5 and 6.6). These sites reflect specific amino acid positions that are likely crucial to the functioning of the various expressed proteins. Codon sites indicated in red, (*mp*=19 sites, *cp*=48 sites and *rep*=104 sites) reflect particular residues within proteins that presently have amino acid states that are broadly adaptive in the context of multiple dicot-infecting mastrevirus niches. Whereas sites in orange (*mp*=2 sites, *cp*=32 sites and *rep*=27 sites) also reflect functionally important amino acid positions, the most adaptive amino acid at these positions differs from species to species. The amino acids at these sites are likely adaptive to features of niches that are specific to the different species. The large number of sites within the various coding regions that appear to be consistently evolving under negative selection is possibly due to the fact that these species all occupy similar ecological niches: something that is unsurprising since they all have largely overlapping host ranges and similar vector species.

There are also a number of interesting patterns in the codon sites that are detectably evolving under either constant (in blue) or episodic (in green) positive selection (i.e. sites at which dN/dS is significantly higher than one in all or a significant fraction of lineages in the particular datasets analysed). For example, the first 52 codon sites of *rep* (the gene region encoding the portion of Rep involved in recognition and binding to the virion strand origin of replication) contains an unusually high proportion of sites (7/52 in CpCDV and 5/52 in CpCV with codon 52 evolving under episodic positive selection in both) that are evolving either under constant (indicated in blue) or episodic (indicated in green) positive selection (i.e. selection favouring change). This suggests that the optimal amino acid configuration in this part of the protein is in a state of flux, with, for example, different configurations perhaps being optimally suited to the different host species that these viruses infect. It is also noteworthy that this is the precise region of *rep* that is either most frequently exchanged

during recombination amongst these viruses, or, when it is transferred, is frequently adaptive and therefore yields genomes that are favoured by natural selection.

In the portion of *rep* encoding the actual origin of replication recognition sequences (called the iteron related domain - labelled IRD in Fig. 6.7), three uniformly spaced codon sites at positions 12, 15 and 18 are detectably evolving under negative selection. This suggests that despite the apparently fluctuating selection pressures acting on the remainder of the DNA binding regions (rolling-circle replication motifs; RCR motifs I, II and III, and the three superfamily helicase motifs; Walker-A, Walker-B and Motif C) of the *rep*, the selective pressures on origin recognition are relatively constant across all the species examined here. Another strikingly conserved pattern of positively and negatively selected codon sites occurs within the region of *rep* encoding the Walker-A motif (Fig. 6.7). This region of *rep* however also falls within the portion of the gene that is expressed in two different frames in Rep and RepA (indicated by a grey shaded box in Fig. 6.7). The apparently conserved positive selection signals detectable throughout this region of overlap between *rep* and *repA* are therefore possibly an artefact of negative selection acting simultaneously on the different proteins these genes encode. For example, the positive selection signals detected in the *rep* codons between positions 215 and 219 could simply be a consequence of negative selection acting to preserve the amino acid coding potential of overlapping codons in *repA*. Regardless of the causes of these positive selection signals the highly conserved negative selection signals at positions 201 and 203 (both encoding a glutamic acid in three species) clearly indicate that selection is strongly favouring these two amino acids within the Rep Walker-A motif.

Other notably conserved negative selection signals occur in *mp* both between codons 6 and 10, and between codons 40 and 81. Whereas conserved negative selection signals are pervasive throughout *cp*, codons 94, 98, and 235 are all detectably evolving under episodic diversifying selection in both TYDV and CpCAV. Although the region encompassing codons 94 and 98 has not been identified as playing any role in transmission the C-terminal region has been associated with vector specificity in the begomoviruses AbMV and TYLCV. Therefore it is possible that codon 235 maybe an important site for vector specificity or transmission efficiency.

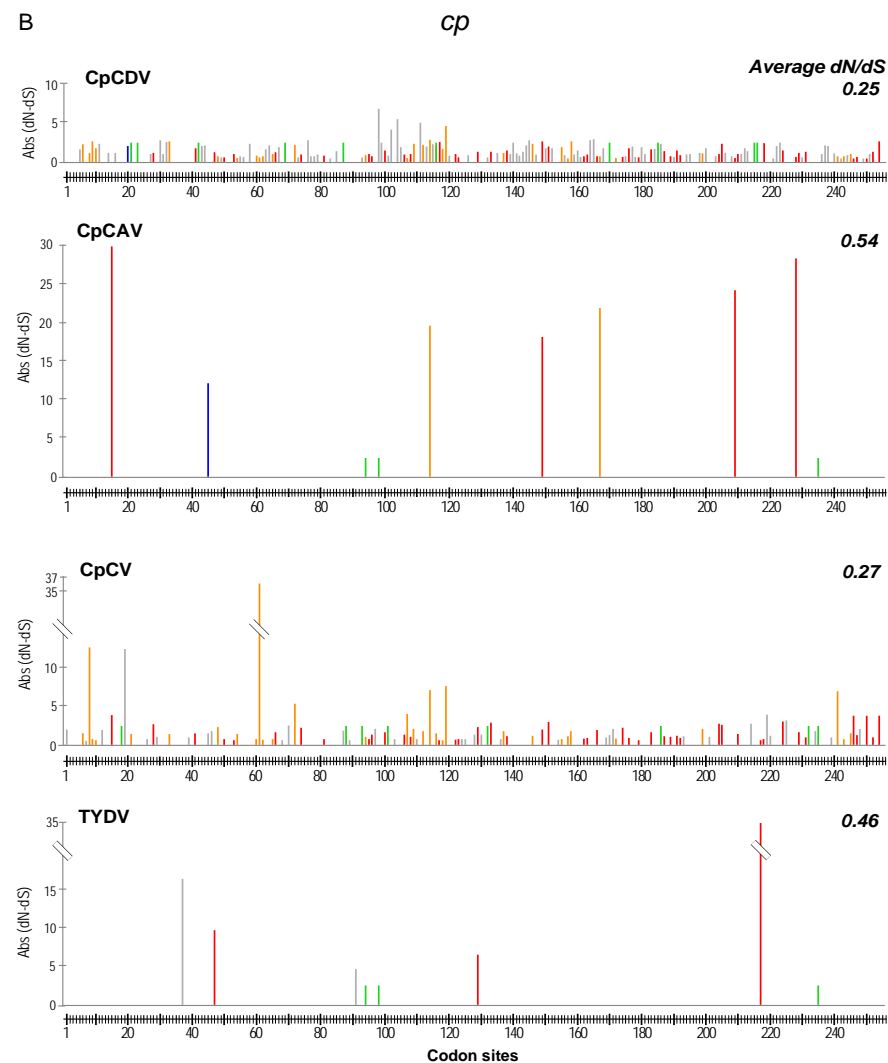
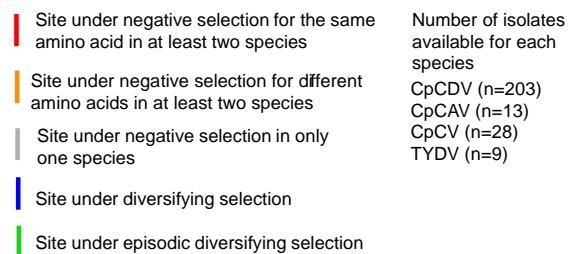
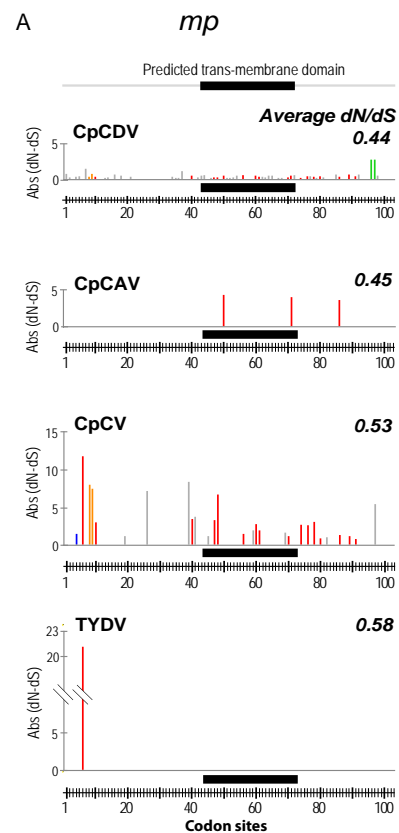


Figure 6.6: See following page for figure legend

Figure 6.6: Plot representing signals of natural selection acting on individual codon sites within **A.** the *mp* and **B.** the *cp* of CpCDV, CpCAV, CpCV and TYDV. Absolute (Abs) values of dN-dS are plotted for positive selection (blue) and negative selection (orange, red and grey) signals with an associated FUBAR p-value <0.05. Abs values for episodic positive selection signals with an associated MEME p-value <0.05 are given in green. Bar heights for Abs (dN-dS) values correspond to the degree of positive or negative selection detected using FUBAR. Sites at which episodic diversifying selection was detected using MEME have been represented by green bars with uniform height across the genes since Abs (dN-dS) values averaged across the entire phylogeny do not accurately reflect degrees of episodic diversifying selection (which by definition occurs only on specific subsets of branches within the phylogeny). Overall, averages for the dN/dS ratios (all of which are significantly <1) are indicated for each gene and species. Codon sites are indicated based on a codon alignment of all species for each gene. The locations of the predicted trans-membrane domain (Boulton *et al.*, 1993) in relation to their position in these alignments are shown. dN=Non-synonymous substitution rates and dS= Synonymous substitution rates.

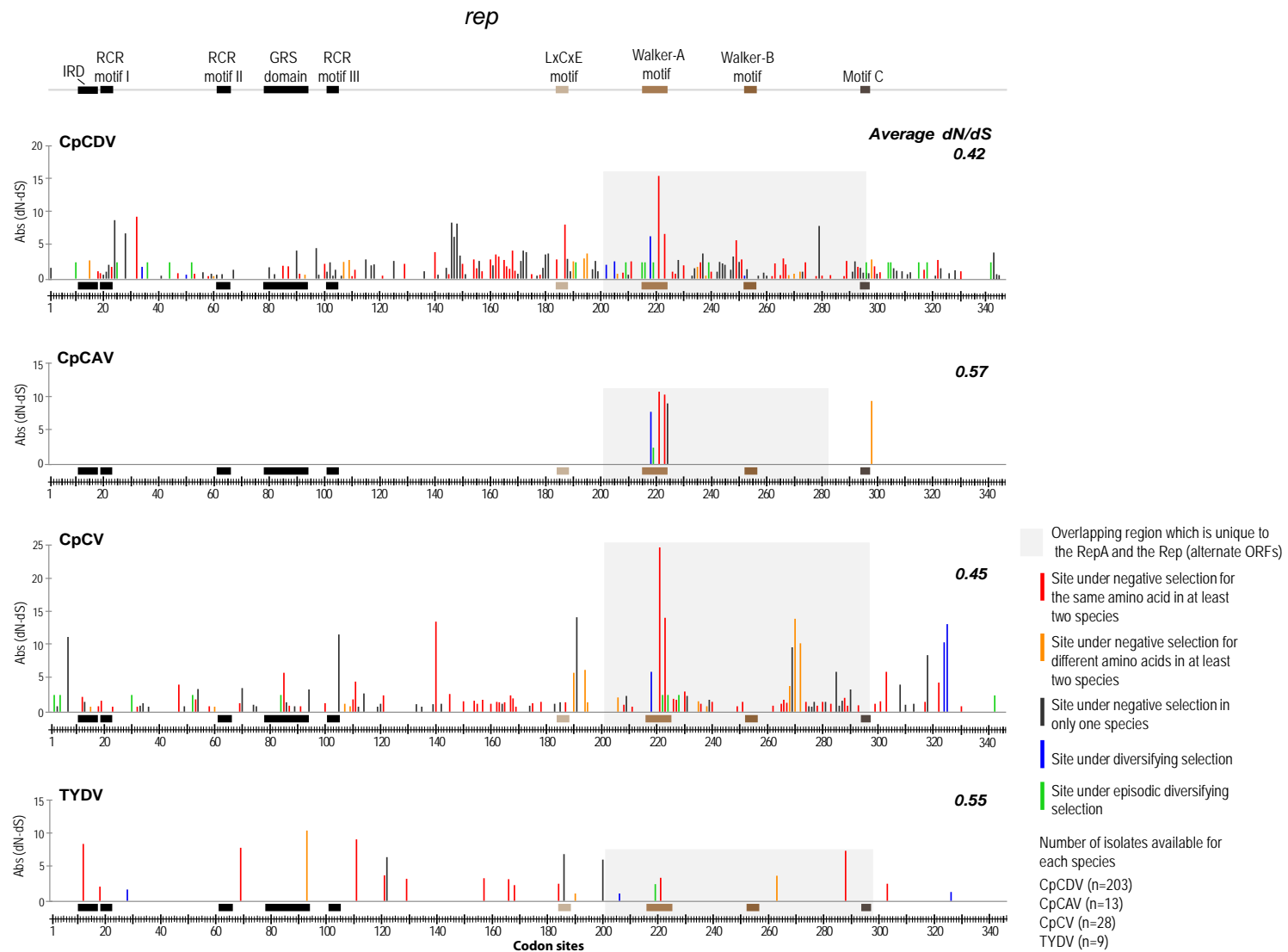


Figure 6.7: See following page for figure legend

Figure 6.7: Plot representing significant signals of natural selection acting on individual codon sites within the *rep* of CpCDV, CpCAV, CpCV and TYDV. Absolute (Abs) values of dN-dS are plotted for positive selection (blue) and negative selection (orange, red and grey) signals with an associated FUBAR p-value <0.05. Abs values for episodic positive selection signals with an associated MEME p-value <0.05 are given in green. Bar heights for Abs (dN-dS) values correspond to the degree of positive or negative selection detected using FUBAR. Sites at which episodic diversifying selection was detected using MEME have been represented by green bars with uniform height across the genes since Abs (dN-dS) values averaged across the entire phylogeny do not accurately reflect degrees of episodic diversifying selection (which by definition occurs only on specific subsets of branches within the phylogeny). Overall, averages for the dN/dS ratios (all of which are significantly <1) are indicated for each gene and species. Codon sites are indicated based on a codon alignment of all species for each gene. The locations of conserved domains and motifs in relation to their positions in these alignments are shown; iteron-related domain (IRD) (Argüello-Astorga & Ruiz-Medrano, 2001), rolling circle replication (RCR) motifs I, II and II (Ilyina & Koonin, 1992; Laufs *et al.*, 1995; Rosario *et al.*, 2012), the geminivirus Rep sequence (GRS) domain (Nash *et al.*, 2011), and the helicase domain Walker-A, -B and -C motifs (Gorbalenya & Koonin, 1993; Gorbalenya *et al.*, 1990). dN=Non-synonymous substitution rates and dS= Synonymous substitution rates.

6.5 Concluding remarks

In this study we analysed the diversity of CpCDV in Sudan by identifying, cloning and sequencing 145 CpCDV genomes from symptomatic pulse samples and included two full genome CpCDV sequences from Sudan which had previously deposited in GenBank. In addition an opportunistically sampled CpCDV isolate from Morocco was recovered. Amongst these isolates four new CpCDV strains have been identified, all with complex patterns of recombination. CpCDV-H, the predominant strain circulating in Sudan, is evidently also frequently recombining with large numbers of other CpCDV strains found within the country. The high frequencies of inter-strain recombination evident within these Sudanese viruses likely reflect high frequencies of infections containing multiple CpCDV strains.

Recently CpCDV has been found in the field infecting previously unsuspected hosts such as cotton and peppers (Akhtar *et al.*, 2013; Manzoor *et al.*, 2014). These and other recent CpCDV diversity studies have raised many questions with regard to the natural host range of this dicot-infecting mastrevirus species, and the role that genetic recombination might play in facilitating its emergence as a pathogen of important crops such as cotton (which is currently one of Sudan's principal cash crops). A survey in Sudan in 2002 using serological tests showed a high incidence of CpCDV-like mastreviruses in various wild plant species, one of which, pigeon pea (*Cajanus cajan*), is commonly planted by Sudanese farmers on the margins of their fields. This has prompted the suggestion that, in this country at least, pigeon pea may facilitate the circulation of CpCDV between other currently unknown uncultivated reservoir species and pulse crops (Ali *et al.*, 2004). In Australia, TYDV-like mastreviruses were shown to have a wide host range (infecting species in seven dicot families) in areas and at a time when chickpeas were not widely grown (Thomas & Bowyer 1984). It is likely that chickpea is particularly susceptible to, and visibly affected by, mastreviruses. In order to enable more informed CpCDV control strategies it might be worthwhile in future CpCDV sampling surveys to attempt the characterisation of both CpCDV variants and other presently undiscovered dicot-infecting mastrevirus species that infect uncultivated dicotyledonous plant species that are commonly found growing in close proximity to important crop species such as cotton and chickpea.

GenBank accession numbers: KM229768 – KM229913

Additional Table 6.1: Genotype and variant designations of all CpCDV isolates, their corresponding GenBank accession numbers and their countries of origin.

CpCDV Strain	Genotype/ Variants	GenBank no.	Country	CpCDV Strain	Genotype/ Variants	GenBank no.	Country	
CpCDV-A	CpCDV-A _{1-I}	FR687959	Syria	CpCDV-H	CpCDV-F _{1-II} CpCDV-F _{2-I}	KC172673	Yemen	
		KC172653	Iran			KF111683	Oman	
		KC172654	Iran			KM229790	Sudan	
		KC172655	Iran			KM229791	Sudan	
		KC172656	Iran		KM229792	Sudan		
		KC172657	Iran		CpCDV-F _{3-I}	KC172674	Eritria	
		KC172658	Iran			KC172675	Eritria	
		KC172659	Iran			CpCDV-H _{1-I}	KC172676	Eritria
		KC172660	Iran			KM229843	Sudan	
		KC172661	Iran		CpCDV-H _{1-II}	KM229798	Sudan	
		KC172662	Turkey			KM229801	Sudan	
		KC172663	Turkey			KM229804	Sudan	
CpCDV-B	CpCDV-B _{1-I}	AM849096	Pakistan	KM229807		Sudan		
		Y11023	South Africa	KM229808	Sudan			
CpCDV-B	CpCDV-B _{2-I}	DQ458791	South Africa	KM229814	Sudan			
	CpCDV-C	CpCDV-C _{1-I}	AM849097	Pakistan	KM229815	Sudan		
AM850136			Pakistan	KM229816	Sudan			
AM900416			Pakistan	KM229817	Sudan			
JF831147			India	KM229818	Sudan			
JF831148			India	KM229821	Sudan			
JX183063			India	KM229822	Sudan			
JX183064			India	KM229823	Sudan			
JX183065			India	KM229828	Sudan			
HG934858			Pakistan	KM229829	Sudan			
KM377673			Pakistan	KM229832	Sudan			
CpCDV-C _{1-II}		KM229768	Sudan	KM229846	Sudan			
		KM229771	Sudan	KM229859	Sudan			
		KM229772	Sudan	KM229862	Sudan			
		KM229773	Sudan	KM229863	Sudan			
		KM229774	Sudan	KM229864	Sudan			
		KM229775	Sudan	KM229870	Sudan			
		KM229776	Sudan	KM229874	Sudan			
		KM229778	Sudan	KM229875	Sudan			
		KM229779	Sudan	KM229876	Sudan			
		KM229782	Sudan	KM229877	Sudan			
		KM229784	Sudan	KM229881	Sudan			
		KM229785	Sudan	KM229883	Sudan			
CpCDV-C _{2-I}		KM229769	Sudan	KM229887	Sudan			
		KM229770	Sudan	KM229893	Sudan			
		KM229777	Sudan	KM229894	Sudan			
		KM229780	Sudan	KM229897	Sudan			
		CpCDV-H _{1-III} CpCDV-H _{1-IV}	KM229781	Sudan	KM229892	Sudan		
			KM229783	Sudan	KM229793	Sudan		
CpCDV-D			CpCDV-D _{1-I}	KM229786	Sudan	KM229794	Sudan	
	FR687960			Pakistan	KM229795	Sudan		
	KC172664	India		KM229796	Sudan			
	KC172665	India		KM229797	Sudan			
	KF176552	India		KM229799	Sudan			
	KF176553	India		KM229800	Sudan			
	KM377672	Pakistan		KM377669	Pakistan			
	KM377670	Pakistan		KM229802	Sudan			
	KM377671	Pakistan		KM229803	Sudan			
	KM229788	Morocco		KM229805	Sudan			
	CpCDV-D _{1-II}	KM229787	Sudan	KM229806	Sudan			
	CpCDV-D _{1-III}	KM377668	Pakistan	KM229809	Sudan			
CpCDV-E	CpCDV-E _{1-I}	AM933134	Sudan	KM229810	Sudan			
		AM933135	Sudan	KM229811	Sudan			
		KM229789	Sudan	KM229812	Sudan			
CpCDV-F	CpCDV-F _{1-I}	KC172666	Pakistan	KM229813	Sudan			
		KC172667	Syria	KM229819	Sudan			
		KC172668	Syria	KM229820	Sudan			
		KC172669	Yemen	KM229824	Sudan			
		KC172670	Yemen	KM229825	Sudan			
		KC172671	Yemen	KM229826	Sudan			
						KM229827	Sudan	

Additional Table 6.1 continued

CpCDV Strain	Genotype/ Variants	GenBank no.	Country	CpCDV Strain	Genotype/ Variants	GenBank no.	Country
CpCDV-H		KC172672	Yemen	CpCDV-L	CpCDV-L _{1-I}	HE864164	Pakistan
		KM229830	Sudan			HE956705	Pakistan
		KM229831	Sudan			HE956706	Pakistan
		KM229833	Sudan			HG313782	Pakistan
		KM229834	Sudan	CpCDV-M	CpCDV-M _{1-I}	KM229908	Sudan
		KM229835	Sudan			KM229909	Sudan
		KM229836	Sudan	CpCDV-N	CpCDV-N _{1-I}	KM229910	Sudan
		KM229837	Sudan	CpCDV-O	CpCDV-O _{1-I}	KM229911	Sudan
		KM229838	Sudan	CpCDV-P	CpCDV-P _{1-I}	KM229912	Sudan
		KM229839	Sudan		CpCDV-P _{1-II}	KM229913	Sudan
		KM229840	Sudan				
		KM229841	Sudan				
		KM229842	Sudan				
		KM229844	Sudan				
		KM229845	Sudan				
		KM229847	Sudan				
		KM229848	Sudan				
		KM229849	Sudan				
		KM229851	Sudan				
		KM229852	Sudan				
		KM229853	Sudan				
		KM229854	Sudan				
		KM229855	Sudan				
		KM229856	Sudan				
		KM229857	Sudan				
		KM229858	Sudan				
		KM229860	Sudan				
		KM229861	Sudan				
		KM229865	Sudan				
		KM229867	Sudan				
		KM229868	Sudan				
		KM229869	Sudan				
		KM229871	Sudan				
		KM229872	Sudan				
		KM229873	Sudan				
		KM229878	Sudan				
		KM229879	Sudan				
		KM229880	Sudan				
		KM229882	Sudan				
		KM229884	Sudan				
		KM229886	Sudan				
		KM229888	Sudan				
		KM229889	Sudan				
		KM229890	Sudan				
		KM229891	Sudan				
		KM229895	Sudan				
		KM229896	Sudan				
		KM229898	Sudan				
		KM229899	Sudan				
	CpCDV-H _{1-V}	KM229850	Sudan				
		KM229885	Sudan				
CpCDV-I	CpCDV-I _{1-I}	KC172677	Eritria				
	CpCDV-I _{2-I}	KM229900	Sudan				
	CpCDV-I _{2-II}	KC172678	Eritria				
CpCDV-K	CpCDV-K _{1-I}	KM229901	Sudan				
		KC172679	Eritria				
		KC172681	Eritria				
		KC172682	Eritria				
	CpCDV-K _{1-II}	KM229902	Sudan				
		KM229903	Sudan				
	CpCDV-K _{1-III}	KC172680	Eritria				
	CpCDV-K _{1-IV}	KM229904	Sudan				
		KM229906	Sudan				
		KM229907	Sudan				
	CpCDV-K _{2-I}	KM229905	Sudan				

6.6 References

- Abraham, A., Menzel, W., Lesemann, D.-E., Varrelmann, M. & Vetten, H. (2006).** Chickpea chlorotic stunt virus: a new polerovirus infecting cool-season food legumes in Ethiopia. *Phytopathology* **96**, 437-446.
- Akhtar, K. P., Ahmad, M., Shah, T. M. & Atta, B. M. (2011).** Transmission of chickpea chlorotic dwarf virus in chickpea by the leafhopper *Orosius albicinctus* (Distant) in Pakistan -short communication. *Plant Protection Science* **47**, 1-4.
- Akhtar, S., Khan, A. J. & Briddon, R. W. (2013).** A Distinct Strain of Chickpea chlorotic dwarf virus Infecting Pepper in Oman. *Plant Disease* **98**, 286-286.
- Ali, M. A., Kumari, S. G., Makkouk, K. H. & Hassan, M. M. (2004).** Chickpea chlorotic dwarf virus, CpCDV naturally infects Phaseolus bean and other wild species in the Gezira region of Sudan. *Arab Journal of Plant Protection* **22**, 96.
- Argüello-Astorga, G. R. & Ruiz-Medrano, R. (2001).** An iteron-related domain is associated to Motif 1 in the replication proteins of geminiviruses: Identification of potential interacting amino acid-base pairs by a comparative approach. *Arch Virol* **146**, 1465-1485.
- Boni, M. F., Posada, D. & Feldman, M. W. (2007).** An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176**, 1035-1047.
- Boulton, M. I., Pallaghy, C. K., Chatani, M., MacFarlane, S. & Davies, J. W. (1993).** Replication of Maize Streak Virus Mutants in Maize Protoplasts: Evidence for a Movement Protein. *Virology* **192**, 85-93.
- Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. (2012).** jModelTest 2: more models, new heuristics and parallel computing. *Nature methods* **9**, 772-772.
- Delpont, W., Poon, A. F. Y., Frost, S. D. W. & Kosakovsky Pond, S. L. (2010).** Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**, 2455-2457.
- Duffy, S. & Holmes, E. C. (2008).** Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. *Journal of Virology* **82**, 957-965.
- Edgar, R. C. (2004).** MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797.
- Farzadfar, S., Pourrahim, R., Golnaraghi, A. R. & Ahoonmanesh, A. (2008).** PCR detection and partial molecular characterization of Chickpea chlorotic dwarf virus in naturally infected sugar beet plants in Iran. *Journal of Plant Pathology* **90**, 247-251.
- Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. (2000).** Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573-582.
- Gorbalenya, A. E. & Koonin, E. V. (1993).** Helicases: amino acid sequence comparisons and structure-function relationships. *Current Opinion in Structural Biology* **3**, 419-429.
- Gorbalenya, A. E., Koonin, E. V. & Wolf, Y. I. (1990).** A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. *FEBS letters* **262**, 145-148.
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010).** New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**, 307-321.
- Hadfield, J., Thomas, J. E., Schwinghamer, M. W., Kraberger, S., Stainton, D., Dayaram, A., Parry, J. N., Pande, D., Martin, D. P. & Varsani, A. (2012).** Molecular characterisation of dicot-infecting mastreviruses from Australia. *Virus Research* **166**, 13-22.

- Hamed, A. A. & Makkouk, K. M. (2002).** Occurrence and management of Chickpea chlorotic dwarf virus in chickpea fields in northern Sudan. *Phytopathologia Mediterranea* **41**, 193-198.
- Horn, N. M., Reddy, S. V., Roberts, I. M. & Reddy, D. V. R. (1993).** Chickpea chlorotic dwarf virus, a new leafhopper-transmitted geminivirus of chickpea in India. *Annals of Applied Biology* **122**, 467-479.
- Ilyina, T. V. & Koonin, E. V. (1992).** Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaeobacteria. *Nucleic Acids Research* **20**, 3279-3285.
- Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. W. (2006).** GARD: a genetic algorithm for recombination detection. *Bioinformatics* **22**, 3096-3098.
- Kraberger, S., Harkins, G. W., Kumari, S. G., Thomas, J. E., Schwinghamer, M. W., Sharman, M., Collings, D. A., Briddon, R. W., Martin, D. P. & Varsani, A. (2013).** Evidence that dicot-infecting mastreviruses are particularly prone to inter-species recombination and have likely been circulating in Australia for longer than in Africa and the Middle East. *Virology* **444**, 282-291.
- Kraberger, S., Mumtaz, H., Claverie, S., Martin, D. P., Briddon, R. W. & Varsani, A. (2014).** Identification of an Australian-like dicot-infecting mastrevirus in Pakistan. *Arch Virol*, In press.
- Kraberger, S., Thomas, J. E., Geering, A. D. W., Dayaram, A., Stainton, D., Hadfield, J., Walters, M., Parmenter, K. S., van Brunschot, S., Collings, D. A., Martin, D. P. & Varsani, A. (2012).** Australian monocot-infecting mastrevirus diversity rivals that in Africa. *Virus Research* **169**, 127-136.
- Kumari, S. G., Makkouk, K. M., Attar, N., Ghulam, W. & Lesemann, D. E. (2004).** First Report of Chickpea chlorotic dwarf virus infecting spring chickpea in Syria. *Plant Disease* **88**, 424-424.
- Kumari, S. G., Makkouk, K. M., Loh, M. H., Negassi, K., Tsegay, S., Kidane, R., Kibret, A. & Tesfatsion, Y. (2008).** Viral diseases affecting chickpea crops in Eritrea. *Phytopathologia Mediterranea* **47**, 42-49.
- Laufs, J., Schumacher, S., Geisler, N., Jupin, I. & Gronenborn, B. (1995).** Identification of the nicking tyrosine of geminivirus Rep protein. *FEBS letters* **377**, 258-262.
- Lefevre, P., Lett, J.-M., Varsani, A. & Martin, D. (2009).** Widely conserved recombination patterns among single-stranded DNA viruses. *Journal of virology* **83**, 2697-2707.
- Liu, L., van Tonder, T., Pietersen, G., Davies, J. W. & Stanley, J. (1997).** Molecular characterization of a subgroup I geminivirus from a legume in South Africa. *Journal of General Virology* **78**, 2113-2117.
- Maddison, W. P. & Maddison, D. R. (2011).** Mesquite: A modular system for evolutionary analysis. Version 2.75 <http://mesquiteproject.org>.
- Makkouk, K., Dafalla, G., Hussein, M. & Kumari, S. (1995).** The natural occurrence of chickpea chlorotic dwarf geminivirus in chickpea and faba bean in the Sudan. *Journal of Phytopathology* **143**, 465-466.
- Makkouk, K. M., Fazlali, Y., Kumari, S. G. & Farzadfar, S. (2002).** First record of Beet western yellows virus, Chickpea chlorotic dwarf virus, Faba bean necrotic yellows virus and Soybean dwarf virus infecting chickpea and lentil crops in Iran. *Plant Pathology* **51**, 387-387.
- Makkouk, K. M., Hamed, A. A., Hussein, M. & Kumari, S. G. (2003).** First report of Faba bean necrotic yellows virus (FBNYV) infecting chickpea (*Cicer arietinum*) and faba bean (*Vicia faba*) crops in Sudan. *Plant Pathology* **52**, 412-412.

- Manzoor, M., Ilyas, M., Shafiq, M., Haider, M., Shahid, A. & Briddon, R. (2014).** A distinct strain of chickpea chlorotic dwarf virus (genus Mastrevirus, family Geminiviridae) identified in cotton plants affected by leaf curl disease. *Arch Virol* **159** (5), 1217-1221.
- Martin, D. & Rybicki, E. (2000).** RDP: Detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562-563.
- Martin, D. P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P. & Varsani, A. (2011a).** Recombination in Eukaryotic Single Stranded DNA Viruses. *Viruses* **3**, 1699-1738.
- Martin, D. P., Briddon, R. W. & Varsani, A. (2011b).** Recombination patterns in dicot-infecting mastreviruses mirror those found in monocot-infecting mastreviruses. *Arch Virol* **156**, 1463-1469.
- Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D. & Lefeuvre, P. (2010).** RDP3: A flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**, 2462-2463.
- Martin, D. P., Posada, D., Crandall, K. A. & Williamson, C. (2005).** A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Research and Human Retroviruses* **21**, 98-102.
- Monjane, A. L., Harkins, G. W., Martin, D. P., Lemey, P., Lefeuvre, P., Shepherd, D. N., Oluwafemi, S., Simuyandi, M., Zinga, I., Komba, E. K., Lakoutene, D. P., Mandakombo, N., Mboukoulida, J., Semballa, S., Tagne, A., Tiendrebeogo, F., Erdmann, J. B., van Antwerpen, T., Owor, B. E., Flett, B., Ramusi, M., Windram, O. P., Syed, R., Lett, J. M., Briddon, R. W., Markham, P. G., Rybicki, E. P. & Varsani, A. (2011).** Reconstructing the history of maize streak virus strain a dispersal to reveal diversification hot spots and its origin in southern Africa. *Journal of Virology* **85**, 9623-9636.
- Morris, B. A. M., Richardson, K. A., Haley, A., Zhan, X. & Thomas, J. E. (1992).** The nucleotide sequence of the infectious cloned dna component of tobacco yellow dwarf virus reveals features of geminiviruses infecting monocotyledonous plants. *Virology* **187**, 633-642.
- Muhire, B., Martin, D. P., Brown, J. K., Navas-Castillo, J., Moriones, E., Zerbini, M. F., Rivera-Bustamante, R. F., Malathi, V. G., Briddon, R. W. & Varsani, A. (2013).** A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). *Arch Virol* **158**, 1411-1424.
- Muhire, B. M., Varsani, A. & Martin, D. P. (2014).** SDT: A Virus Classification Tool Based on Pairwise Sequence Alignment and Identity Calculation. *PLoS ONE* **9**, e108277.
- Mumtaz, H., Kumari, S. G., Mansoor, S., Martin, D. P. & Briddon, R. W. (2011).** Analysis of the sequence of a dicot-infecting mastrevirus (family *Geminiviridae*) originating from Syria. *Virus Genes* **42**, 422-428.
- Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Pond, S. L. K. & Scheffler, K. (2013).** FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Molecular biology and evolution* **30**, 1196-1205.
- Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K. & Pond, S. L. K. (2012).** Detecting individual sites subject to episodic diversifying selection. *Plos Genet* **8**, e1002764.
- Nahid, N., Amin, I., Mansoor, S., Rybicki, E., van der Walt, E. & Briddon, R. (2008).** Two dicot-infecting mastreviruses (family *Geminiviridae*) occur in Pakistan. *Arch Virol* **153**, 1441-1451.
- Nash, T. E., Dallas, M. B., Reyes, M. I., Buhrman, G. K., Ascencio-Ibañez, J. T. & Hanley-Bowdoin, L. (2011).** Functional analysis of a novel motif conserved across geminivirus Rep proteins. *Journal of Virology* **85**, 1182-1192.

- Owor, B. E., Shepherd, D. N., Taylor, N. J., Edema, R., Monjane, A. L., Thomson, J. A., Martin, D. P. & Varsani, A. (2007). Successful application of FTA[®] Classic Card technology and use of bacteriophage ϕ 29 DNA polymerase for large-scale field sampling and cloning of complete maize streak virus genomes. *Journal of Virological Methods* **140**, 100-105.
- Padidam, M., Sawyer, S. & Fauquet, C. M. (1999). Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218-225.
- Posada, D. & Crandall, K. A. (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 13757-13762.
- Rosario, K., Duffy, S. & Breitbart, M. (2012). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol* **157**, 1851-1871.
- Shackelton, L. A., Parrish, C. R., Truyen, U. & Holmes, E. C. (2005). High rate of viral evolution associated with the emergence of carnivore parvovirus. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 379-384.
- Shepherd, D. N., Martin, D. P., Lefevre, P., Monjane, A. L., Owor, B. E., Rybicki, E. P. & Varsani, A. (2008). A protocol for the rapid isolation of full geminivirus genomes from dried plant tissue. *Journal of Virological Methods* **149**, 97-102.
- Smith, J. M. (1992). Analyzing the mosaic structure of genes. *Journal of Molecular Evolution* **34**, 126-129.
- Stenzel, T., Piasecki, T., Chrzęstek, K., Julian, L., Muhire, B. M., Golden, M., Martin, D. P. & Varsani, A. (2014). Pigeon circoviruses display patterns of recombination, genomic secondary structure and selection similar to those of beak and feather disease viruses. *Journal of General Virology* **95**, 1338-1351.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**, 2713-2739.
- Thomas, J., Parry, J., Schwinghamer, M. & Dann, E. (2010). Two novel mastreviruses from chickpea (*Cicer arietinum*) in Australia. *Arch Virol* **155**, 1777-1788.
- Varsani, A., Monjane, A. L., Donaldson, L., Oluwafemi, S., Zinga, I., Komba, E. K., Plakoutene, D., Mandakombo, N., Mboukoulida, J., Semballa, S., Briddon, R. W., Markham, P. G., Lett, J. M., Lefevre, P., Rybicki, E. P. & Martin, D. P. (2009). Comparative analysis of Panicum streak virus and Maize streak virus diversity, recombination patterns and phylogeography. *Virology Journal* **6**, 194.
- Varsani, A., Shepherd, D. N., Monjane, A. L., Owor, B. E., Erdmann, J. B., Rybicki, E. P., Peterschmitt, M., Briddon, R. W., Markham, P. G., Oluwafemi, S., Windram, O. P., Lefevre, P., Lett, J. M. & Martin, D. P. (2008). Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. *Journal of General Virology* **89**, 2063-2074.

Chapter 7

Identification of novel circular DNA viruses associated with *Poaceae* species in New Zealand

Contents

7.1	Abstract.....	240
7.2	Introduction.....	241
7.3	Materials and Methods	242
7.3.1	Sample collection and DNA extraction	242
7.3.2	Next-generation sequencing and virus isolation.....	244
7.3.3	Full genome analysis and pairwise comparison	244
7.3.4	Phylogenetic analysis.....	244
7.4	Results and discussion	245
7.4.1	Novel viral genome analysis	245
7.4.2	Rep and CP analysis.....	248
7.4.3	Identification of putative motifs in the Replication-associated protein and iterons in the long intergenic region	252
7.5	Concluding remarks.....	255
7.6	References.....	256

This body of work has been published in Archives of Virology and is presented in a similar manner to that of the publication:

Krabberger, S., Farkas, K., Bernardo, P., Booker, C., Argüello-Astorga, G. R., Mesléard, F., Martin, D. P., Varsani, A. (2015) Identification of novel Bromus- and Trifolium-associated circular DNA viruses. Archives of Virology. DOI 10.1007/s00705-015-2358-6

7.1 Abstract

The genomes of a large number of highly diverse novel circular DNA viruses from a wide range of sources have been characterised in recent years. Some of these recovered circular single-stranded DNA (ssDNA) viruses share similarities to the plant-infecting ssDNA virus family, *Geminiviridae*. Here we describe four novel circular ssDNA viral genomes which encode replication-associated proteins (Rep) that are most closely related to those of either geminiviruses or gemycircularviruses (a potentially new family of ssDNA viruses that is closely related to geminiviruses). These four viral genomes were recovered from *Bromus hordeaceus* in New Zealand. Two of these viral genomes share >99% and have tentatively been named Bromus-associated circular virus-1 (BasCV-1), and the two other divergent genomes have tentatively been named BasCV-2 and BasCV-3. BasCV-1 shares ~57% identity to geminivirus-like, Nepavirus (previously isolated from sewage; GenBank accession number JQ898333), BasCV-2 shares ~57% identity with SaCV-3 (also from sewage; GenBank accession number KJ547627) and BasCV-3 shares between 56% and 64% sequence identity to, and phylogenetically clusters with, members of the gemycircularvirus group. All four of the viral genomes recovered have a major open-reading frame on both their complementary and virion sense strands, one of which likely encodes a Rep and the other a coat protein. Although future infectivity studies are needed to identify the host(s) of these viruses, this is the first report of ssDNA viruses that are associated with grasses in New Zealand.

7.2 Introduction

The previous Chapters collectively entail a comprehensive look into the overall dynamics of mastreviruses in regions of the world where these viruses are known to exist. Given geminiviruses are found globally and Australia harbours a high level of mastrevirus diversity it is entirely plausible that mastreviruses or related viruses may be present in New Zealand. The aim of this endeavour was to take an exploratory approach using a viral metagenomic method to investigate the existence of mastreviruses or related viruses in wild grasses in New Zealand.

Poaceae is one of the largest plant families and contains more than 10,000 species of grasses which collectively have a near global distribution (Barker *et al.*, 2001). Although members of this family are infected by a variety of viruses, the only single-stranded DNA viruses that are known to infect members of the *Poaceae* family are mastreviruses (family: *Geminiviridae*). Of these, the most well studied is *Maize streak virus* (MSV) because of its devastating impact on maize production in Africa (Martin *et al.*, 1999; Monjane *et al.*, 2011; Shepherd *et al.*, 2010; Varsani *et al.*, 2008). MSV together with other species of monocot-infecting mastreviruses infect both cultivated and non-cultivated grasses and have been identified in Africa, Europe, Asia and Australia (Candresse *et al.*, 2014). A recent study has revealed a high degree mastrevirus diversity within uncultivated Australian grasses (Krabberger *et al.*, 2012). Based on the wide distribution of mastreviruses and the close proximity of New Zealand to Australia, a mastrevirus diversity hotspot, it is plausible that a similar degree of mastrevirus diversity might also occur within New Zealand's grasses.

Several grass-infecting viruses have been documented in New Zealand, many of which are thought to have been introduced into New Zealand through the movement of exotic grass species (Davis & Guy, 2001). There have been twelve identified species, all of which are RNA viruses: *Barley yellow dwarf virus-PAV*, *Barley yellow dwarf virus-RMV*, *Barley yellow dwarf virus-MAV*, *Barley stripe mosaic virus*, *Cocksfoot mottle virus*, *Cocksfoot mild mosaic virus*, *Cynosurus mottle virus*, *Cereal yellow dwarf virus-RPV*, *Ryegrass cryptic virus*, *Rye grass mosaic virus*, *Soil-borne wheat mosaic virus* and *Wheat streak mosaic virus*.

(Delmiglio *et al.*, 2010; Guy, 2006; 2014; Latch, 1977; Mohamed, 1978; Pearson *et al.*, 2006).

Recent advances in sequencing technologies have enabled the application of sequence independent metagenomic approaches in the detection and characterisation of entirely novel viral communities that are associated with various different eukaryote species or environments (Dayaram *et al.*, 2014; McDaniel *et al.*, 2013; Ng *et al.*, 2011; Rosario *et al.*, 2009; Sikorski *et al.*, 2013; van den Brand *et al.*, 2012). These methods have been particularly effective at exploring the diversity within the environment of diverse circular replication-associated protein (Rep) encoding ssDNA (CRESS-DNA) viruses. Consequently this has yielded the discovery of both highly divergent members of known ssDNA virus families (Candresse *et al.*, 2014; Kreuze *et al.*, 2009; Victoria *et al.*, 2009) and CRESS-DNA viruses that potentially belong to entirely new families: such as the recently proposed gemycircularvirus family (Rosario *et al.*, 2012a; Sikorski *et al.*, 2013). In an attempt to identify and characterise ssDNA viruses associated with *Poaceae* sp. in New Zealand we firstly used such an approach to identify viruses associated with 33 uncultivated grass samples, and then recovered full genomes and Sanger sequenced four CRESS-DNA genomes from four of these samples.

7.3 Materials and Methods

7.3.1 Sample collection and DNA extraction

Grasses presenting symptoms commonly associated with virus-infections (foliar yellowing and crinkling) were opportunistically collected from locations in the North (n=7) and South island (n=26) of New Zealand between 2012 and 2014. A total of 13 *Poaceae* species were sampled, *Anthoxanthum odoratum* (n=3), *Arrhenatherum elatius* (n=5), *Axonopus fissifolius* (n=1), *Bromus hordeaceus* (n=4), *Bromus inermis* (n=5), *Bromus tectorum* (n=1), *Dactylis glomerata* (n=3), *Elymus farctus* (n=1), *Holcus lanatus* (n=5), *Hordeum vulgare* (n=1), *Paspalum dilatatum* (n=1), *Pennisetum clandestinum* (n=1) and *Poa alpine* (n=2).

Leaf material was homogenised in 1ml of SM Buffer [0.1 M NaCl, 50 mM Tris-HCl (pH 7.4)]. The homogenate was centrifuged at 10,000 x g for 10 min to pellet cellular debris and the supernatant was passed through 0.45µm and 0.2µm syringe filters in succession (Sartorius Stedim Biotech, Germany). Viral DNA was isolated using the High Pure Viral Nucleic Acid Kit (Roche Diagnostics, USA). Circular viral DNA in the DNA extract was enriched using TempliPhiTM (GE Healthcare, USA). The enriched DNA was then sequenced on an Illumina HiSeq 2000 sequencer at the Beijing Genomics Institute (Hong Kong).

7.3.2 Next-generation sequencing and virus isolation

Paired-end reads were assembled using ABySS 1.3.6 (Simpson *et al.*, 2009) with a k-mer setting of 64. The assembled contigs >500nt were analysed for significant similarities to proteins of geminiviruses and geminivirus-like CRESS-DNA viruses (indicated by BLASTx E-scores $<10^{-5}$). Back-to-back primers were designed based on conserved regions within the *rep* of geminivirus-like viral contigs in order to recover full viral genomes from individual grass samples using polymerase chain reaction (PCR). PCR was performed using specific back-to-back primer pairs (Table 1) and Kapa HiFi Hotstart DNA polymerase (Kapa Biosystems, USA) with the following thermocycler conditions: 94°C for 3 min, 25 cycles of 98 °C (20sec), 60 °C (30sec), 72 °C (3min) and a final extension of 72°C for 3 min. The resulting amplicons of ~2.5–2.7 kb were gel-purified using the Quick-spin PCR Product Purification Kit (iNtRON Biotechnology, Korea) and purified products were ligated into pJET1.2 (Thermo Fisher Scientific Inc, USA). The resulting clones were Sanger sequenced at Macrogen Inc. (Korea) by primer walking. The Sanger sequence contigs were assembled into full genome sequences using DNA Baser V4 (Heracle BioSoft S.R.L. Romania).

7.3.3 Full genome analysis and pairwise comparison

Full genome sequences were managed using MEGA5 (Tamura *et al.*, 2011) and open reading frames identified using DNA man software (version 5.2.9; Lynnon Biosoft) and BLASTx (Altschul *et al.*, 1990). Pairwise comparison of those genomes recovered in this study together with those most similar available in GenBank, was undertaken using SDT V1.2 software (Muhire *et al.*, 2014).

7.3.4 Phylogenetic analysis

A Rep amino acid sequence dataset was compiled consisting of the sequences from the four CRESS-DNA viruses recovered in this study, representative sequences from the gemycircularviruses, geminiviruses and nanoviruses. Additionally, the Reps of BamiV, NimiV and NephV and geminivirus-like Rep sequences from fungal and rhizaria genomes were included. We aligned these sequences using MUSCLE (Edgar, 2004) and constructed a maximum likelihood (ML) phylogenetic tree using PHYML version 3 (Guindon *et al.*, 2010) with the substitution model RtRev+G chosen using ProtTest 3 (Darriba *et al.*, 2011) and the

approximate likelihood-ratio test (aLRT) used to estimate branch support (Anisimova & Gascuel, 2006). The nanovirus Rep sequences were used to root the tree.

7.4 Results and discussion

7.4.1 Novel viral genome analysis

Four probable CRESS-DNA viral genomes were recovered from leaf material of *Bromus hordeaceus* collected in two locations in New Zealand; Sefton, Canterbury (n=2) and Mt Victoria, Wellington (n=2) (Table 7.1). BLASTx analysis of these four viruses revealed that they each likely encode a Rep which is similar to those of gemycircularviruses and geminiviruses. These genomes have at least one major ORF in both the virion and complementary sense. One of the first gemycircularvirus to be characterised was the fungal-infecting virus *Sclerotinia sclerotiorum* hypovirulence-associated DNA virus-1 (SsHADV-1) (Yu *et al.*, 2010). Although SsHADV-1 was first identified in China, it has subsequently been found in both benthic sediments in New Zealand rivers and associated with dragonflies and damselflies in the USA (Dayaram *et al.*; Krabberger *et al.*, 2013). Other viruses that are closely related to SsHADV-1 have been recovered from variety of other sources. These include two viral genomes, one from cassava (Cassava-associated circular DNA virus; CasCV) and another from *Hypericum japonicum* leaf material (Hypericum japonicum-associated circular DNA virus; HjasCV) (Dayaram *et al.*, 2012; Du *et al.*, 2014). Other gemycircularvirus genomes have been isolated from faecal matter (Meles meles fecal virus; MmFV and Faecal-associated gemycircularviruses; FaGmV-1–13) (Ng *et al.*, 2014; Sikorski *et al.*, 2013; van den Brand *et al.*, 2012), insects (Mosquito VEM SDBVL-G virus; MvemV and Dragonfly-associated circular viruses; DfasCV-1–5) (Dayaram *et al.*; Ng *et al.*, 2011; Rosario *et al.*, 2012a), healthy bovine serum (HCBI8 and HCBI9) (Lamberto *et al.*, 2014) and serum from a patient with multiple sclerosis (MSSI2) (Lamberto *et al.*, 2014).

Seven CRESS-DNA viral genomes that are distantly related to geminiviruses but do not group with the gemycircularviruses (Baminivirus; BamiV, Niminivirus; NimiV, Nephavirus; NephV and SaCV-1–4 (Ng *et al.*, 2012) have previously been found in either treated or raw sewage material. Additionally four distantly related genomes have been recovered from insect species in the order Odonata (Odonata-associated circular DNA virus; OdasCV-6, -7, -

8, -9 and -15) (Dayaram *et al.*) and one from ancient caribou faeces (Ancient caribou feces associated virus; anCFV (Ng *et al.*). Interestingly, Rep-like sequences similar to the geminiviruses have also been identified in various fungal and rhizaria genomes (Liu *et al.*, 2011).

Table 7.1: Details for Bromus-associated circular DNA viral isolates and back-to-back primers used to recover full genomes.

Genbank accession #	Virus name	Isolate ID	New Zealand location	Forward primer	Reverse primer
KM510189	BasCV-1	NZG01-29-Sef	Sefton, Canterbury	GAACAGCTGAACGAGGGTTT	CAGTGGAGATGTAGCTTCGA
KM510190	BasCV-1	NZG03-29-Wel	Mt Victoria, Wellington	GAACAGCTGAACGAGGGTTT	CAGTGGAGATGTAGCTTCGA
KM510191	BasCV-2	NZG03-39-Wel	Mt Victoria, Wellington	GCCTGCTTCACGATACAC	GAGACCGCTTCTAGTGCT
KM510192	BasCV-3	NZG01-118-Sef	Sefton, Canterbury	AGAACGTAATCCCCGGTG	TTGGAGGAGACCTTGACG

NB: BasCV=Bromus-associated circular DNA virus

Based on genome-wide pairwise identities, the four probable viral genomes recovered here from *Bromus hordeaceus* are highly diverged from previously characterised CRESS-DNA viruses and most likely can be classified as three distinct viral species which we have tentatively named Bromus-associated circular DNA virus (BasCV) -1, -2 and -3. The genomes of BasCV-1, -2 and -3 are 2776 nt, 2542 nt and 2238 nt in size, respectively (Fig. 7.1).

Genome-wide pairwise comparisons using SDT 1.2 (Muhire *et al.*, 2014) indicated that the two isolates NZG03-29 (Mt Victoria) and NZG01-29 (Sefton) share ~99% pairwise identity with one another and ~57% with Nepavirus (JQ898333). We have tentatively named these two viruses BasCV-1 [NZ-NZG01-Sef-2012] (KM510189) and BasCV-1 [NZ-NZG03-Wel-2012] (KM510190). Isolate NZG03-39 (Mt Victoria) which shares ~57% pairwise identity to SaCV-3 (KJ547627) have tentatively been named BasCV-2 [NZ-NZG03-Wel-2012] (KM510191). Lastly, isolate NZG01-118 (Sefton) which shares <64% genome-wide pairwise identity with all other currently known gemycircularviruses has been tentatively named BasCV-3 [NZ-NZG01-Sef-2012] (KM510192). Comparison of the three BasCVs with each other indicated that while they all share 55-58% genome-wide identity with one another, they all have different nonanucleotide sequences as their probable virion strand origins of replication: TATATAA(A/G)A (BasCV-1), TAGTATTAC (BasCV-2) and TAATGTTAT (BasCV-3).

7.4.2 Rep and CP analysis

Within the four viral genomes we identified putative Reps and capsid proteins (CPs). In addition to the two major open reading frames (ORFs), in BasCV-1 and -2 we identified two other minor ORFs (Fig. 7.1A and B). In BasCV-1 and -3 we identified putative introns in the Rep and therefore the Reps may potentially be expressed from spliced transcripts as seen in members of the geminivirus genus, *Mastrevirus* (Dekker *et al.*, 1991) (Fig. 7.1A and C). The Rep of BasCV-1 shares ~32% amino acid identity with that of NepaV (Fig. 7.1A), whereas the BasCV-3 Rep shares ~63% with those of DfaCV-2 and CasCV (Fig. 7.1C). Similar to the *repA* gene found in mastreviruses, putative ORFs encoding a *repA* were also identified in BasCV-1 and 3. The Rep of BasCV-2 is also likely encoded in the complementary sense and shares 34% identity with the Rep of Rodent-stool circovirus M-45 (JF755409) and McMurdo ice shelf pond associated circular DNA virus-2 (KJ547647) (Fig. 7.1B). The putative CP

ORF of BasCV-3 shares significant levels of similarity to known CRESS viruses including ~41% amino acid identity to the CP amino acid sequences of gemycircularviruses (Fig. 7.1C).

A Rep amino acid maximum likelihood tree was constructed which included the four CRESS-DNA viruses recovered in this study, representative sequences from the gemycircularviruses, geminiviruses and nanoviruses. Also included were the Repls and Rep-like sequences of BamiV, NimiV and NephV and geminivirus-like Rep sequences from fungal and rhizaria genomes. Due to the high degree of sequence variation amongst the geminivirus-like sequences we were only able to credibly align, and therefore use, these Rep amino acid sequences to infer their phylogenetic relationships (Fig. 7.2). The phylogenetic tree shows that the Repls of the BasCVs are indeed distantly related to those of the geminiviruses. The Rep of BasCV-3 is nested in the monophyletic gemycircularvirus clade whereas the Repls of BasCV-1 and BaCV-2 branch basal to both the geminiviruses and geminivirus groupings.

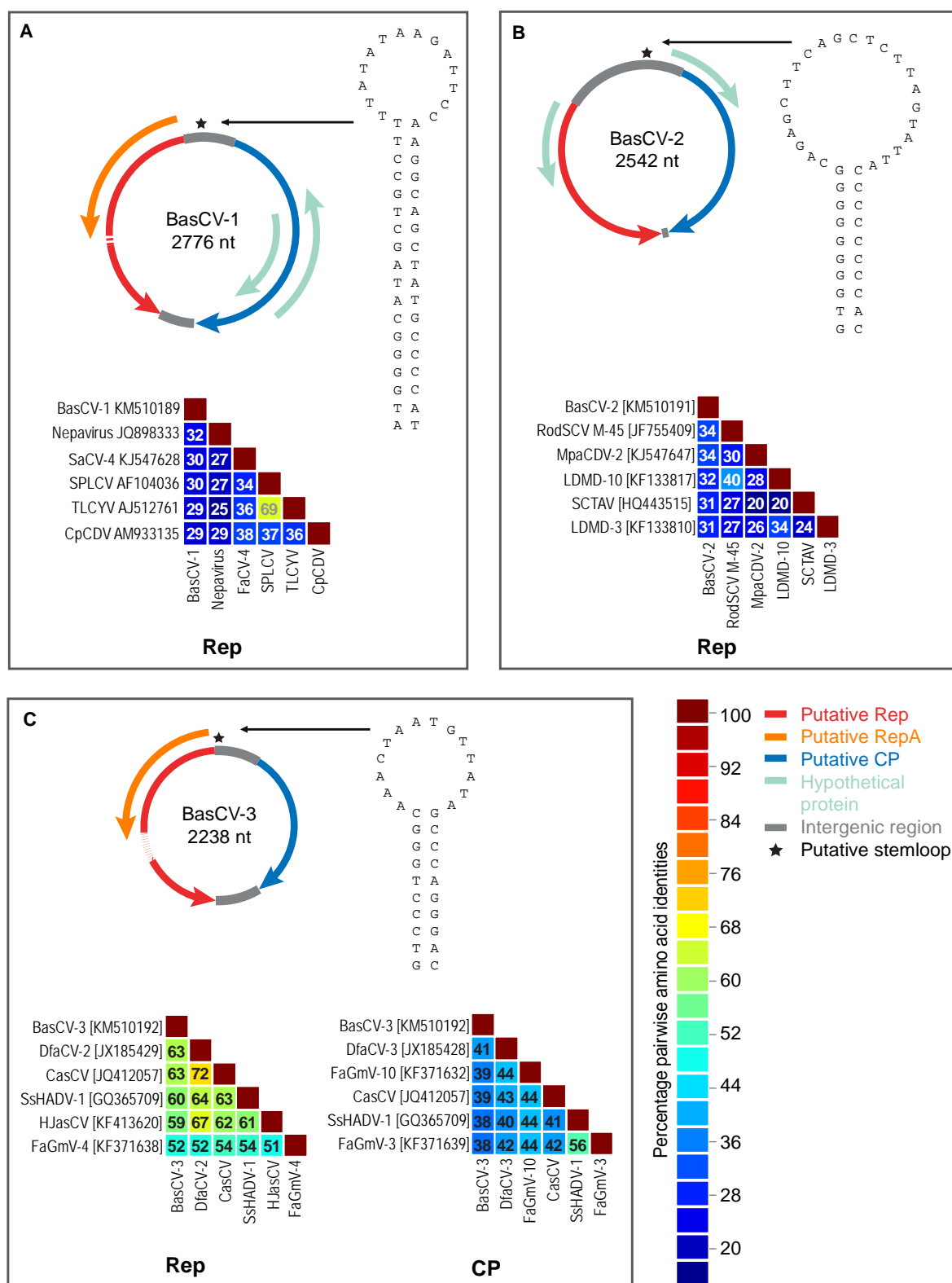


Figure 7.1: Genome organisation of Bromus-associated circular DNA virus -1, 2 and 3. Stemloop structure with nonanucleotide sequence containing the probable virion strand origin of replication (A). Percentage pairwise amino acid identities of BasCV-1, -2 and -3 (five highest identities) of the Rep (B) and CP (C).

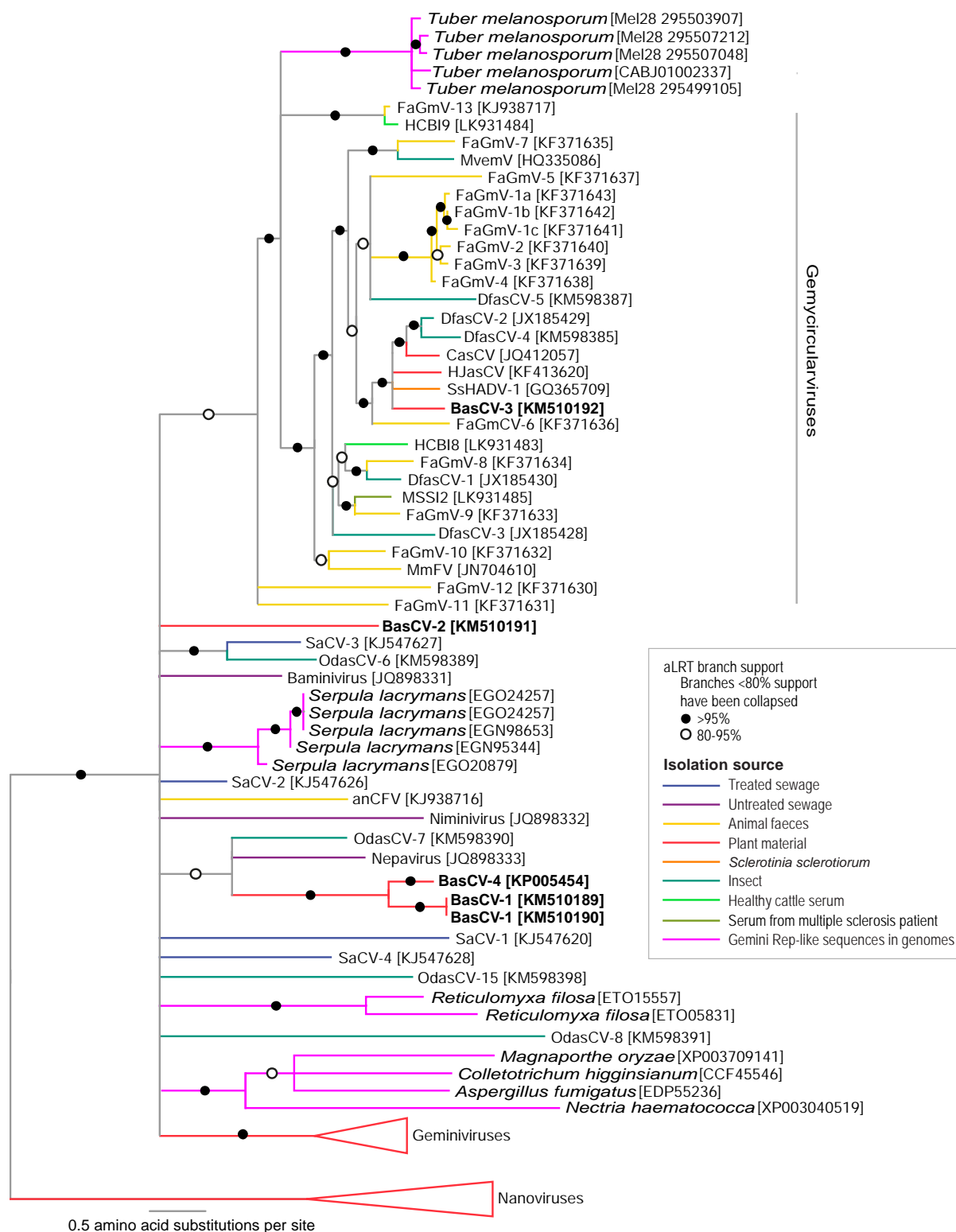


Figure 7.2: Maximum-likelihood phylogenetic tree of the Rep amino acid sequences (rooted with Reps of nanoviruses) of BasCV-1, 2 and 3, gemycircularviruses, other geminivirus-like CRESS viruses, geminiviruses and Rep-like sequences in fungal genomes. Branches with aLRT branch support of <80% have been collapsed. Sample names in bold represent those recovered in this study.

7.4.3 Identification of putative motifs in the Replication-associated protein and iterons in the long intergenic region

Within the Reps encoded by BasCVs, we identified both the three rolling circle replication related motifs, and the three superfamily 3 (SF3) helicase motifs which are commonly found in CRESS-DNA viruses (Ilyina & Koonin, 1992; Koonin, 1993; Koonin & Ilyina, 1992; Rosario *et al.*, 2012b) (Fig. 7.3). The Rep motifs identified for BasCV-1 and -3 are more similar to geminivirus RCR motifs whereas those identified in BasCV-2 are more similar to those found in circoviruses. A conserved region of Rep found in geminiviruses known as the geminivirus Rep sequence (GRS) domain (Fig. 7.3) (Nash *et al.*, 2011) which is thought to be involved in cleavage of ssDNA during RCR, was identified in BasCV-1 and BasCV-3. We note that BasCV-2 does not seem to have a GRS domain. The Walker motif A, B and motif C (SF3 helicase motifs) seem to be conserved across the Reps of BasCV isolates and most CRESS-DNA viruses. Various *cis* acting elements have been identified in the intergenic regions of ssDNA viruses which are thought to interact with the Rep to initiate RCR (Argüello-Astorga *et al.*, 1994; Dayaram *et al.*, 2014; Dayaram *et al.*, 2013; Londoño *et al.*, 2010). The specificity of these repeating sequence elements (called iterons in geminiviruses) has been shown to play an important part in the efficiency of replication (Herrera-Valencia *et al.*, 2006). We identified potential analogues of these elements in the BasCV-1 and BasCV-3 sequences (Fig. 7.4). These elements are similar to those found in some of the gemycircularviruses, BasCV-3 which groups with the gemycircularviruses shares the same iteron sequence with the ssHADV-1 which is known to infect fungi (Fig. 7.4B). This gives weight to the idea that BasCV-3 may in fact be hosted by an endo- or epiphytic fungus of this bromus grass.

	Motif I	Motif II	GRS domain	Motif III	Walker A	Walker B	Motif C
BasCV-1 [KM510190]	LLTYAQC	NPHIHAVE	YDDTCESYHFNISPAQ	CINYCRGKNK	VVIIGPSGSGKT	IVFDEV	FTCT
BasCV-2 [KM510191]	CFTSFSE	REHWQGYAE	WQCELSHNAHIECRK	AIEYCRKEET	LVLVGPSRLGKT	LIFDDI	VLCN
BasCV-3 [KM510192]	LLTYAQC	GVHLHCFVD	DVFDVEGRHFNISPSK	GYDYAIKDGR	ICVYGESRTGKT	AVFDDI	WLSN
DfaCV-1 [JX185430]	LLTYPQC	GVHLHAFPM	RVFDVDGHHFNIVRGY	GWAYATKDGD	LILIGDTRLGKT	AVFDDM	YISN
DfasCV-2 [JX185429]	LVTYPQC	GLHLHCFAD	DIFDVGDCHPNIQPS	GYDYAIKDGD	LVLVGESRTGKT	AIFDDI	WISN
DasfCV-3 [JX185428]	LLTYAQS	GTHYHAFPM	RIFDIDGYPHNILSGR	MYDYATKDGD	LVLVGPSRTGKT	AVFDDI	WCNN
CasCV [JO412057]	LITYAQC	GVHLHCFID	DIFDVGDRHFNIEPSW	GYDYAIKDGD	LVLVGDSRSRGKT	AIFDDI	WISN
SsHADV-1 [GQ365709]	LLTYAQC	GTHLHCFAE	DVFDVDGHHFNITKSR	GYDYAIKDGD	LVLVGPSQTGKT	AVFDDI	WCSN
HJasCV [KF413620]	LVTYAQC	GLHLHVAFD	DILDVDGRHFNLPAPK	AYDYAIKDGD	LVLFGGTRTGKT	AVFDDI	WICN
MmFV [JN704610]	LLTYAQC	GIHLHAFVD	RRFDVEGFHFNIPCG	MLDYAIKDGD	LILWGETRLGKT	AVLDDM	WLMN
MvemV [HQ335086]	LLTYAQC	GIHFHAFLD	RFWDIAGRHPNIARVG	AYDYAIKDND	LVLFGPSRTGK-	AVFDDI	WVSN
FaGmV-1 [KF371643]	LLTYAQC	GTHLHAFVD	DVFDVGGRHFNLPVSY	GFDYAIKDGD	LVIYGDTRLGKT	AVFDDM	WLAN
FaGmV-2 [KF371640]	LLTYAQC	GTHLHAFCD	DVFDVGGRHFNVMPSF	GWYATKDGD	LVIYGDTRLGKT	AVFDDM	WLSN
FaGmV-3 [KF371639]	LLTYAQC	GTHLHAFCD	DVFDVGGRHFNLPVSY	GYDYAIKDGD	LVIYGDTRLGKT	AVFDDM	WLSN
FaGmV-4 [KF371638]	LLTYAQC	GTHLHAFCD	DVFDVGGFHFNIEASR	GYDYAIKDGD	LVIYGDTRLGKT	AVFDDM	WLAN
FaGmV-5 [KF371637]	LVTYPQS	GTHLHVCFD	DIFDVGGFHFNIESK	GALYACKDGD	LVLIGDALTGKT	GVIDDI	WIAN
FaGmV-6 [KF371636]	LLTYAQC	GIHLHCFAD	RIFDVGDRHFNVPVSR	GYDYAIKDGD	LILYGPSLTGKT	AVLDDI	WCAN
FaGmV-7 [KF371635]	LLTYPQI	GYTSHCFLD	RIFDIQGHFNIEVRG	AYNYTIKDND	LVMYGDSRLGKT	AIFDDI	WCSN
FaGmV-8 [KF371634]	LLTYPQS	GLHLHAFPM	RVFDVDGRHFNIVRGY	GATYAIKDGD	LVLVGPTRLGKT	AIFDDM	YICN
FaGmV-9 [KF371633]	LLTYAQC	GIHLHAFVD	RVFDVQGHFNIEVSR	GYDYAIKDGD	LVLVGPTRTGKT	AVFDDF	WINN
FaGmV-10 [KF371632]	SVTLCPH	GTHLHAFCD	RRFDVDGYPHNVPQFG	GWYAIKDGD	LVLVGESRLGKT	AVFDDM	WLCN
FaGmV-11 [KF371631]	FLTYSQY	GHYHVFLVA	RIFDVGDCHPNFKSVR	LPGYCLDGD	LCLIGRSRLGKT	AVMDDI	WCTN
FaGmV-12 [KF371630]	FLTYSQV	GFHFHAFIL	RIFDFDGLHFNIESVR	KITYTKKDGE	SQMLGPHRRRT	AVFDDI	WVCN
MSSI2 [LK931485]	LLTYPQC	GLHLHAFVD	RAFDVEGCHPNVSPSR	GFDYAIKDGD	LVVYGPSRMGKT	AIFDDF	WLSN
HCB19 [LK931484]	IITFPQV	GIHYHIYLG	TAFDYFAGHGNIKSIR	VFDYVGKDGD	LILWGPTRTGKT	AVFDDI	MCMN
HCB18 [LK931483]	LLTYAQC	GTHLHAFVD	AVFDVGGFHPNISI	HYDYAIKDGD	LVLNGASRLGKT	AVFDDI	WISN
FaCV-1 [KJ547620]	IITYPQS	SLHRHAYVL	-FFDHLTRHPNICKVG	VIAYVKKCGE	LVLNGASRLGKT	IVFDDI	TSN
FaCV-2 [KJ547626]	LITYPQC	GLHVHAIIV	RFDDVAGFHPNIQTVR	AYTYLDKEPV	LVLIGPSRTGKT	IIFDDV	WLCI
FaCV-3 [KJ547627]	SFTYPQC	GLHLHAYLH	DADFVDGFPNIEQKPR	VIAYCKEDT	PPLLSSPCLYWL	IVFDDI	WLCN
FaCV-4 [KJ547628]	FLTYPQC	SPHLHAYVC	TFFNENYHFNIEQSA	VIAYTKKDGD	LILIGPSKLKGT	IVFDDF	YCAN
Nepavirus [JO898333]	AITYSRC	SFHRHAYVR	EYFKFRHAQCNIQPCR	WNNYVRKDGD	NVLVGPTGCGKT	ILFDDM	ITCN
Nimivirus [JO898332]	FLTYPQC	HPHFHAYLC	RHFDISGYHFNIEQVCR	VLKYVTKDGE	LWLFGPSKTGKS	LVLDDI	VCSN
Baminiavirus [JO898331]	LLTYPQC	NPHLHILWE	RFFDVTDFHPNVVVVR	ARDYIAKTGA	LYVEGASRIGKT	AVFDDI	FLVN
DfasCV-5 [KM598387]	VLLTYSQ	GTHFHVFD	NVFDVGGHFNILPVW	AFDYAAKDGD	HLHGSSLAFRKP	AVFDDW	WLCN
DfasCV-4 [KM598385]	FLITYAQ	GLHLHVFD	DIFDVGDRHFNIEKRS	GYDYAIKDGE	PTLEGGNGSGQT	AIFDDI	WISN
OdasCV-15 [KM598398]	FFLTYPQ	NFHLHALVT	TFFDLNGFHPNIAAK	LKNYISKEDI	VWVKGPSGIGKT	IMLDDM	WLCN
OdasCV-7 [KM598390]	VFLTYSH	TDHFHAVLC	RLFDENGWHPKIESPR	SITYVVKDGD	LILVGPSGCGKT	IIFDDM	FLSN
OdasCV-8 [KM598391]	FFLTYPK	TPHLHVYMQ	RWADLGKYGFFYTTR	VMAYVMKDGN	HMIIGIPDTGKS	VIYNDP	FTCN
OdasCV-6 [KM598389]	VFLTYPQ	QPHIHAYAA	GCFDVGDRHFNIEQKPR	VAEYCGKHDT	LLLVGASKLGKT	IVLDDF	FICN
ACMV [GO204107]	FLTFPKC	QPHLHMLIQ	RFFDLVSPTRSAHFHP	VKSVIDKDGD	IVIEGDSRTGKT	NVIDDV	FLCN
TYLCSV [GU951759]	FLTYPKG	EPHLHALIQ	RLFDVHPSCSTSFHP	VKSYLKDGD	IVIEGDSRTGKT	NVIDDV	FLCN
ECSV [FJ66563]	FLTYSQC	NNHLHAIVC	RIFDFGEFHPKIECTCR	SLKYIQKEAG	LIEGDSRTGKT	NVIDDV	WLCN
TCTV [X84735]	FLTYPHC	EPHLHVLIQ	RHFDLRDGGSGRICH	VKSYLEKDGD	IIIEGPSRTGKT	NVIDDV	ILCN
MSV [X01633]	FLTYPKC	SLHLHALLQ	RFFDINGFHPNIEQSAK	VRDYILKEPL	LYIVGPTRTGKS	NIVDDI	ILAN
BCTIV [JX082259]	FLTYSQI	GFHTHCIIQ	KKLDVNGNLFFNIILP	AFEYITKEDT	LYICGPSRTGKT	HIIDDI	AITN
ABTV [EF546807]	MFTINNP	TRHVQGYVE	RAL--I-----PGAH-	ARAYCMKADT	IWVYGPNGGEGK	IVIFDI	VMAN
BBTV [EU531473]	MFTINNP	TRHVQGYVE	RGF--F-----PGAH-	ARSYCMKEDT	IWVYGPNGGEGK	IVIFDI	VMAN
BFDV [AF071878]	CFTLNPP	TPHLQGYFH	KKMLP-----RG	NEKYCSKEGD	DVIYGPNGGEGK	VILDDF	IITS
BarCV [GU799606]	CFTLNPP	TPHLQGFNL	KKW-----FNARA	NDEYCTKGGD	KVYVGDGCSKS	VIVDDF	YVTS
FaGmV-13 [KJ938717]	IITFPQV	GIHYHVYLG	TAFDYFAGHGNIKSVR	VFDYVGKDGD	LILWGPTRTGKT	AVFDDI	MCMN
AnCFV [KJ938716]	FLTYARA	GIHYHVLC	DSFDVLGHFNWTPIR	EDQHTVKNHK	LLLVGPTRLGKT	LILDDF	WLCQ
R. filosa [ETO15557]	FLTYPQC	GRHLHCYIN	HRDLDDGFGHNYQGR	VKKYCTKEKN	LWLHGKNTGKT	IVFDDM	FTSN
T. melanosporum [CABJ01002337]	LVTYPRC	GIHYHALVH	RCLDVEYGGRVYHPNL	VRDYCRKEGW	LILYGPTRTGKT	VVFDV	GCFN
S. lacrymans [EGO24257]	LLTYSQI	GSHWHIVK	AFDVGDIHPNVKTLA	AWTYLQKEEG	LVVWGPTRTGKT	IVLDDI	---

Figure 7.3: Alignment of Rep sequences showing rolling circle replication motifs I, II and III, Walker motif A and B and motif C of CRESS-DNA viruses (BasCV-1, 2 and 3 - shown in red, the gemycircularviruses, other geminivirus-like CRESS viruses and representative geminiviruses, circoviruses, nanoviruses and Rep-like sequences in fungal and rhizaria genome).

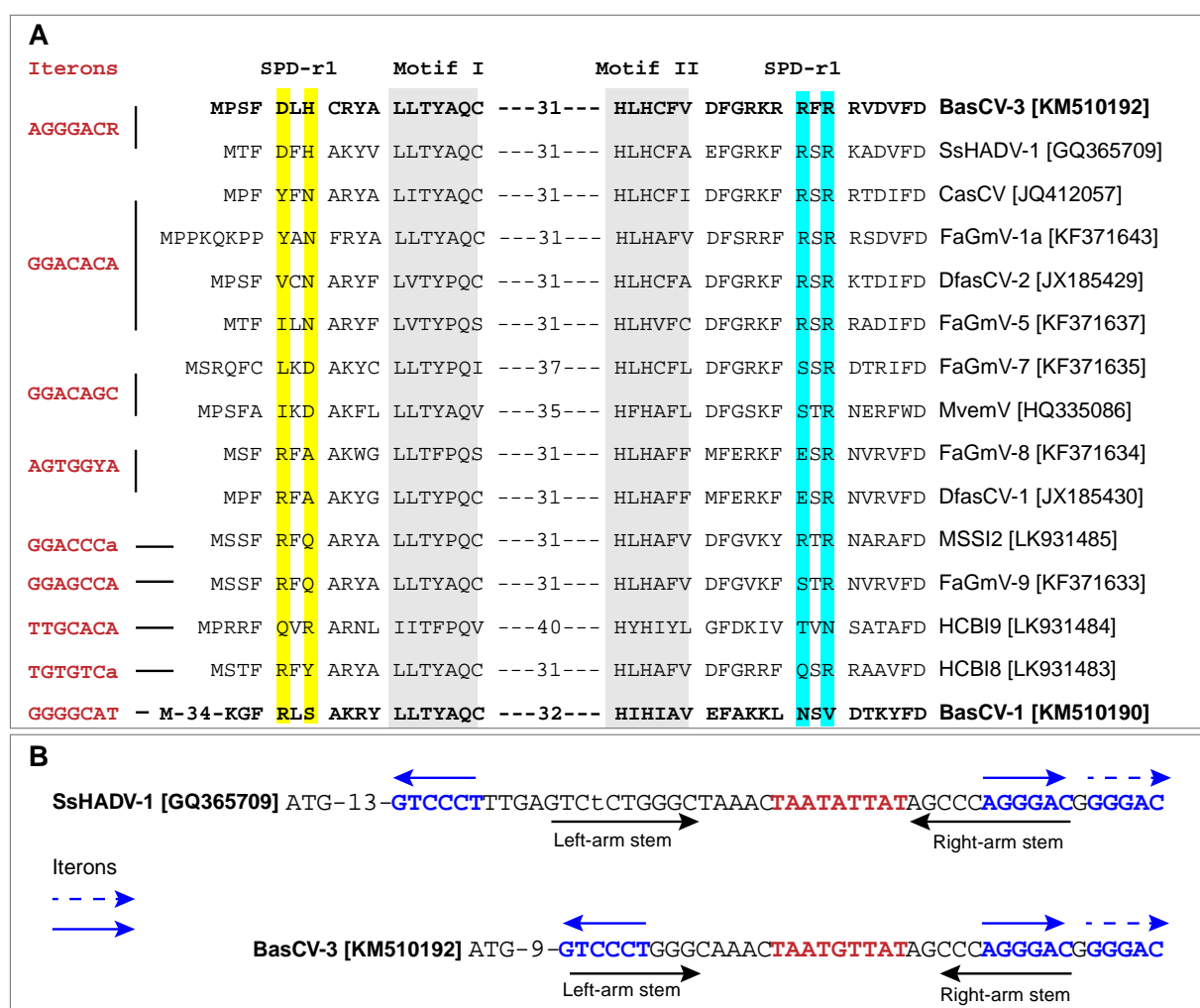


Figure 7.4: (A) Correlations between iterons core sequences and potential Rep DNA-binding SPDs of BasCV-1, BasCV-3 and selected gemycircularviruses. Amino acid residues identified as putative SPDs in the beta-1 strand (r1) are shaded in yellow, whereas SPDs in the beta-strand (r-2) associated to Motif II are shaded in blue. The conserved RCR motifs I and II are indicated at the top of the alignments. The Rep N-end of BasCV-1 is also aligned to show the similitude of its RCR motifs with gemycircularviruses. **(B)** Comparisons of the origin of replication region in the LIR of BasCV-3 and SsHADV-1, illustrating the resemblances between their putative Rep-binding sites (iterons).

7.5 Concluding remarks

In summary, we have found four viral genomes associated with *Bromus hordeaceus*; this is the first report of circular ssDNA viruses being associated with grasses in New Zealand. Two of these, belonging to the tentative species BasCV-1, were found in two geographically distinct locations, one in the South Island and the other in the North Island of New Zealand suggesting that this tentative species has a reasonably wide distribution in New Zealand. We were, however unable to ascertain whether these viruses infect *B. hordeaceus*, or infect other organisms associated with *B. hordeaceus* (e.g. fungi, bacteria or protists). However, it is evident that the Reps encoded by these viruses are most closely related to those of plant-infecting geminiviruses and probably fungus infecting gemycircularviruses. A third genome, tentatively classified as belonging to another new species, BasCV-3, is most closely related to SsHADV-1 (a definitively fungus-infecting gemycircularvirus) and CasCV and HJasCV (viruses previously isolated from cassava and *H. japonicum*, respectively). It is therefore plausible that BasCV-3 may infect one or more Bromus-associated fungus species. Recently gemycircularvirus isolates have also been recovered from healthy cattle serum. This finding, together with the discovery here of similar viruses associated with grasses (a primary component in the diet of most cattle), increases the plausibility that gemycircularviruses are able to move between species in different kingdoms (much as some plant viruses are circulatively transmitted by their insect vectors). An extensive survey of grasses in New Zealand as well as infectivity studies will be necessary to determine the diversity and host range of these viruses.

GenBank accession numbers: KM510189 – KM510192

7.6 References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.
- Anisimova, M. & Gascuel, O. (2006). Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Systematic Biology* **55**, 539-552.
- Argüello-Astorga, G. R., Guevara-González, R. G., Herrera-Estrella, L. R. & Rivera-Bustamante, R. F. (1994). Geminivirus Replication Origins Have a Group-Specific Organization of Iterative Elements: A Model for Replication. *Virology* **203**, 90-100.
- Barker, N. P., Clark, L. G., Davis, J. I., Duvall, M. R., Guala, G. F., Hsiao, C., Kellogg, E. A. & Linder, H. P. (2001). Phylogeny and Subfamilial Classification of the Grasses (Poaceae). *Annals of the Missouri Botanical Garden* **88**, 373-457.
- Candresse, T., Filloux, D., Muhire, B., Julian, C., Galzi, S., Fort, G., Bernardo, P., Daugrois, J.-H., Fernandez, E., Martin, D. P., Varsani, A. & Roumagnac, P. (2014). Appearances Can Be Deceptive: Revealing a Hidden Viral Infection with Deep Sequencing in a Plant Quarantine Context. *PLoS ONE* **9**, e102945.
- Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164-1165.
- Davis, L. T. & Guy, P. L. (2001). Introduced Plant Viruses and the Invasion of a Native Grass Flora. *Biological Invasions* **3**, 89-95.
- Dayaram, A., Galatowitsch, M., Harding, J. S., Argüello-Astorga, G. R. & Varsani, A. (2014). Novel circular DNA viruses identified in *Procordulia grayi* and *Xanthocnemis zealandica* larvae using metagenomic approaches. *Infection, Genetics and Evolution* **22**, 134-141.
- Dayaram, A., Goldstien, S., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2013). Novel single stranded DNA virus recovered from estuarine Mollusc (*Amphibola crenata*) whose replication associated protein (Rep) shares similarities with Rep-like sequences of bacterial origin. *Journal of General Virology* **94**, 1104-1110.
- Dayaram, A., Opong, A., Jäschke, A., Hadfield, J., Baschiera, M., Dobson, R. C. J., Offei, S. K., Shepherd, D. N., Martin, D. P. & Varsani, A. (2012). Molecular characterisation of a novel cassava associated circular ssDNA virus. *Virus Research* **166**, 130-135.
- Dayaram, A., Potter, K. A., Pailles, R., Moline, A. B., Marinov, M., Rosenstein, D. D. & Varsani, A. Identification of diverse circular Rep-encoding DNA viruses in dragonflies and damselflies of Arizona and Oklahoma. In review.
- Dekker, E. L., Woolston, C. J., Xue, Y., Cox, B. & Mullineaux, P. M. (1991). Transcript mapping reveals different expression strategies for the bicistronic RNAs of the geminivirus wheat dwarf virus. *Nucleic Acids Research* **19**, 4075-4081.
- Delmiglio, C., Pearson, M. N., Lister, R. A. & Guy, P. L. (2010). Incidence of cereal and pasture viruses in New Zealand's native grasses. *Annals of Applied Biology* **157**, 25-36.

- Du, Z., Tang, Y., Zhang, S., She, X., Lan, G., Varsani, A. & He, Z. (2014).** Identification and molecular characterization of a single-stranded circular DNA virus with similarities to *Sclerotinia sclerotiorum* hypovirulence-associated DNA virus 1. *Arch Virol* **159**, 1527-1531.
- Edgar, R. C. (2004).** MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797.
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010).** New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**, 307-321.
- Guy, P. L. (2006).** New Zealand grasslands revisited: identification of Cocksfoot mild mosaic virus. *Australasian Plant Pathology* **35**, 461-464.
- Guy, P. L. (2014).** Viruses of New Zealand pasture grasses and legumes: a review. *Crop and Pasture Science* **65**, 841-853.
- Herrera-Valencia, V. A., Dugdale, B., Harding, R. M. & Dale, J. L. (2006).** An iterated sequence in the genome of Banana bunchy top virus is essential for efficient replication. *Journal of General Virology* **87**, 3409-3412.
- Ilyina, T. V. & Koonin, E. V. (1992).** Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. *Nucleic Acids Research* **20**, 3279-3285.
- Koonin, E. V. (1993).** A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the initiation of eukaryotic DNA replication. *Nucleic Acids Research* **21**, 2541-2547.
- Koonin, E. V. & Ilyina, T. V. (1992).** Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. *The Journal of General Virology* **73** 2763-2766.
- Kraberger, S., Stainton, D., Dayaram, A., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2013).** Discovery of *Sclerotinia sclerotiorum* Hypovirulence-Associated Virus-1 in Urban River Sediments of Heathcote and Styx Rivers in Christchurch City, New Zealand. *Genome Announcements* **1**, e00559-00513.
- Kraberger, S., Thomas, J. E., Geering, A. D. W., Dayaram, A., Stainton, D., Hadfield, J., Walters, M., Parmenter, K. S., van Brunschot, S., Collings, D. A., Martin, D. P. & Varsani, A. (2012).** Australian monocot-infecting mastrevirus diversity rivals that in Africa. *Virus Research* **169**, 127-136.
- Kreuze, J. F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I. & Simon, R. (2009).** Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. *Virology* **388**, 1-7.
- Lamberto, I., Gunst, K., Müller, H., zur Hausen, H. & de Villiers, E.-M. (2014).** Mycovirus-Like DNA Virus Sequences from Cattle Serum and Human Brain and Serum Samples from Multiple Sclerosis Patients. *Genome Announcements* **2**.
- Latch, G. (1977).** Incidence of barley yellow dwarf virus in ryegrass pastures in New Zealand. *New Zealand journal of agricultural research* **20**, 87-89.

- Liu, H., Fu, Y., Li, B., Yu, X., Xie, J., Cheng, J., Ghabrial, S. A., Li, G., Yi, X. & Jiang, D. (2011).** Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC Evolutionary Biology* **11**, 276.
- Londoño, A., Riego-Ruiz, L. & Argüello-Astorga, G. (2010).** DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. *Arch Virol* **155**, 1033-1046.
- Martin, D., Willment, J. & Rybicki, E. (1999).** Evaluation of maize streak virus pathogenicity in differentially resistant *Zea mays* genotypes. *Phytopathology* **89**, 695-700.
- McDaniel, L. D., Rosario, K., Breitbart, M. & Paul, J. H. (2013).** Comparative metagenomics: Natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. *Environmental microbiology* **16**, 570-585.
- Mohamed, N. (1978).** Cynosurus mottle virus, a virus affecting grasses in New Zealand. *New Zealand journal of agricultural research* **21**, 709-714.
- Monjane, A. L., Harkins, G. W., Martin, D. P., Lemey, P., Lefevre, P., Shepherd, D. N., Oluwafemi, S., Simuyandi, M., Zinga, I., Komba, E. K., Lakoutene, D. P., Mandakombo, N., Mboukoulida, J., Semballa, S., Tagne, A., Tiendrebeogo, F., Erdmann, J. B., van Antwerpen, T., Owor, B. E., Flett, B., Ramusi, M., Windram, O. P., Syed, R., Lett, J. M., Briddon, R. W., Markham, P. G., Rybicki, E. P. & Varsani, A. (2011).** Reconstructing the history of maize streak virus strain a dispersal to reveal diversification hot spots and its origin in southern Africa. *Journal of Virology* **85**, 9623-9636.
- Muhire, B. M., Varsani, A. & Martin, D. P. (2014).** SDT: A Virus Classification Tool Based on Pairwise Sequence Alignment and Identity Calculation. *PLoS ONE* **9**, e108277.
- Nash, T. E., Dallas, M. B., Reyes, M. I., Buhrman, G. K., Ascencio-Ibañez, J. T. & Hanley-Bowdoin, L. (2011).** Functional analysis of a novel motif conserved across geminivirus Rep proteins. *Journal of Virology* **85**, 1182-1192.
- Ng, T. F. F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., Oderinde, B. S., Wommack, K. E. & Delwart, E. (2012).** High Variety of Known and New RNA and DNA Viruses of Diverse Origins in Untreated Sewage. *Journal of Virology* **86**, 12161-12175.
- Ng, T. F. F., Willner, D. L., Lim, Y. W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y., Rohwer, F. & Breitbart, M. (2011).** Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PLoS ONE* **6**, e20579.
- Ng, T. F. F., Zhou, Y., Chen, L.-F., Shapiro, B., Stiller, M., Heintzman, P. D., Varsani, A., Kondov, N. O., Wong, W., Deng, X., Andrews, T. D., Moorman, B. J., Meulendyk, T., MacKay, G., Gilbertson, R. & Delwart, E. (2014).** Preservation of viral genomes in 700-year-old caribou feces from an subarctic ice patch. *PNAS*, doi:10.1073/pnas.1410429111.
- Pearson, M. N., Clover, G. R. G., Guy, P. L., Fletcher, J. D. & Beever, R. E. (2006).** A review of the plant virus, viroid and mollicute records for New Zealand. *Australasian Plant Pathology* **35**, 217-252.

- Rosario, K., Dayaram, A., Marinov, M., Ware, J., Krabberger, S., Stainton, D., Breitbart, M. & Varsani, A. (2012a). Diverse circular single-stranded DNA viruses discovered in dragonflies (Odonata: Epiprocta). *Journal of General Virology* **93**, 2668-2681.
- Rosario, K., Duffy, S. & Breitbart, M. (2012b). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol* **157**, 1851-1871.
- Rosario, K., Nilsson, C., Lim, Y. W., Ruan, Y. & Breitbart, M. (2009). Metagenomic analysis of viruses in reclaimed water. *Environmental Microbiology* **11**, 2806-2820.
- Shepherd, D. N., Martin, D. P., Van Der Walt, E., Dent, K., Varsani, A. & Rybicki, E. P. (2010). Maize streak virus: An old and complex 'emerging' pathogen. *Molecular Plant Pathology* **11**, 1-12.
- Sikorski, A., Massaro, M., Krabberger, S., Young, L. M., Smalley, D., Martin, D. P. & Varsani, A. (2013). Novel myco-like DNA viruses discovered in the faecal matter of various animals. *Virus Research* **177**, 209-216.
- Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M. & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research* **19**, 1117-1123.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**, 2713-2739.
- van den Brand, J. M. A., van Leeuwen, M., Schapendonk, C. M., Simon, J. H., Haagmans, B. L., Osterhaus, A. D. M. E. & Smits, S. L. (2012). Metagenomic Analysis of the Viral Flora of Pine Marten and European Badger Feces. *Journal of Virology* **86**, 2360-2365.
- Varsani, A., Shepherd, D. N., Monjane, A. L., Owor, B. E., Erdmann, J. B., Rybicki, E. P., Peterschmitt, M., Briddon, R. W., Markham, P. G., Oluwafemi, S., Windram, O. P., Lefeuvre, P., Lett, J. M. & Martin, D. P. (2008). Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. *Journal of General Virology* **89**, 2063-2074.
- Victoria, J. G., Kapoor, A., Li, L., Blinkova, O., Slikas, B., Wang, C., Naeem, A., Zaidi, S. & Delwart, E. (2009). Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *Journal of virology* **83**, 4642-4651.
- Yu, X., Li, B., Fu, Y., Jiang, D., Ghabrial, S. A., Li, G., Peng, Y., Xie, J., Cheng, J., Huang, J. & Yi, X. (2010). A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. *Proceedings of the National Academy of Sciences* **107**, 8387-8392.

Chapter 8

Characterisation of a diverse range of Rep-encoding ssDNA viruses recovered from a sewage treatment oxidation pond

Contents

8.1	Abstract.....	261
8.2	Introduction.....	262
8.3	Materials and methods.....	265
8.3.1	Sample collection and viral DNA isolation.....	265
8.3.2	Next-generation sequencing-informed recovery of complete viral genomes.....	265
8.3.3	Phylogenetic analyses.....	269
8.3.4	Recombination analysis of gemycircularviruses.....	269
8.4	Results and discussion.....	270
8.4.1	Recovery and characterisation of novel CRESS DNA viral genomes.....	270
8.4.1.1	Circular viral genomes recovered by PCR using abutting primers.....	270
8.4.1.2	Subgenomic DNA molecules.....	272
8.4.2	Phylogenetic and sequence analyses of novel sewage-associated circular viruses.....	279
8.4.3	Conserved motifs within replication-associated proteins.....	286
8.4.4	Nonanucleotide sequence analysis.....	289
8.4.5	Recombination patterns among Gemycircularviruses.....	289
8.5	Concluding remarks.....	292
8.6	References.....	296

This body of work has been published in *Infection, Genetics and Evolution*, and is presented in a similar manner to that of the publication:

Kraberger, S., Argüello-Astorga, G. R., Greenfield, L. G., Galilee, C., Law, D., Martin, D. P., Varsani, A., (2015) Characterisation of a diverse range of Rep-encoding ssDNA viruses recovered from a sewage treatment oxidation pond. *Infection, Genetics and Evolution* 31, 73-86.

8.1 Abstract

Our knowledge of circular single-stranded DNA (ssDNA) virus diversity has increased dramatically in recent years, largely due to advances in high-throughput sequencing technologies. These viruses are apparently major virome components in most terrestrial and aquatic environments and it is therefore of interest to determine their diversity at the interfaces between these environments. Treated sewage water is a particularly interesting interface between terrestrial and aquatic viromes in that it is directly pumped into waterways and is likely to contain virus populations that have been strongly impacted by humans. We used a combination of high-throughput sequencing, full genome PCR amplification, cloning and Sanger sequencing to investigate the diversity of circular ssDNA viruses present in a sewage oxidation pond. Using this approach, we recovered 50 putatively complete novel circular ssDNA viral genomes (it remains possible that some are components of multipartite viral genomes) and 11 putatively sub-genome-length circular DNA molecules which may be either defective genomes or components of multipartite genomes. Thirteen of the genomes have bidirectional genome organisations and share similar conserved replication-associated protein (Rep) motifs to those of the gemycircularviruses: A group that in turn is most closely related to the geminiviruses. The remaining 37 viral genomes share very low degrees of Rep similarity to all other known Rep-encoding ssDNA viruses. This number of highly divergent ssDNA virus genomes within a single sewage treatment pond further reinforces the notion that there likely exists hundreds of completely unknown genus/family level ssDNA virus groupings.

8.2 Introduction

The discovery of three novel CRESS DNA viruses associated with wild grasses in New Zealand was discussed in Chapter Seven. These viruses are distantly related to Geminiviruses and two of these viruses share a similar architecture to mastreviruses, including a putative spliced Rep. This is the first time a CRESS DNA virus associated with any grass species in New Zealand has been documented and further validates such sequence independent approaches for the discovery of novel CRESS DNA viruses. The study discussed in this chapter used a similar viral metagenomic approach as in Chapter Seven in an effort to investigate the presence of geminiviruses or related CRESS DNA viruses in treated sewage material and further elucidate the known viral diversity of these viruses in New Zealand.

Sewage is a biological sink for a wide variety of infectious agents including viruses. Consisting largely of human excrement, sewage harbours a rich diversity of viruses. Besides viruses that infect the multitude of environmental microbes which degrade faeces, sewage also contains viruses infecting humans, their gut flora, and their food (Blinkova *et al.*, 2009; Cantalupo *et al.*, 2011; Metcalf *et al.*, 1995; Ng *et al.*, 2012; Parsley *et al.*, 2010; Rosario *et al.*, 2009c; Symonds *et al.*, 2009; Tamaki *et al.*, 2012). Secondary stages of sewage treatment involve aeration in ‘oxidation ponds’ which might also introduce viruses that infect birds, algae, fungi, insects and aerobic bacteria into the treated sewage water that is ultimately discharged into the natural environment.

To date virus research on sewage systems has predominantly focused on viruses of clinical importance to humans, such as poliovirus and other enteroviruses (Blomqvist *et al.*, 2004; Hewitt *et al.*, 2011; Katayama *et al.*, 2008; Lodder & de Roda Husman, 2005; Symonds *et al.*, 2009; Vaidya *et al.*, 2002). Only a handful of studies have taken an unbiased metagenomics-based approach to study viral diversity associated with raw sewerage (Cantalupo *et al.*, 2011; Ng *et al.*, 2012; Tamaki *et al.*, 2012) and one which looked at reclaimed water (Rosario *et al.*, 2009b). These revealed that sewage contains genetic material derived from a wide variety of vertebrate, invertebrate and plant-infecting viruses. A significant proportion of this genetic material encodes proteins that share low, but nevertheless significant, degrees of similarity to those encoded by circular single-stranded

DNA (ssDNA) viruses in the families *Geminiviridae*, *Circoviridae* and *Nanoviridae*. For example, Ng *et al.* (2012) recorded as much as 30% of viral-related sequences assembled from their metagenomic analysis of raw sewage samples shared significant similarities to proteins of these three ssDNA families. If this genetic material is indeed viral-derived, many of these represent completely novel ssDNA virus groups. For example, three complete ssDNA genomes that were recently isolated from raw sewage (named Nepavirus, Nimivirus and Baminivirus) encode replication-associated proteins (Reps) that are most closely related to but clearly distinct from Rep proteins expressed by plant-infecting geminiviruses and each likely represents an undescribed genus / group (Ng *et al.*, 2012). Similarly divergent ssDNA viruses have been found in faecal samples from humans (Castrignano *et al.*, 2013), caribou (Ng *et al.*), sheep (Sikorski *et al.*, 2013b), New Zealand fur seals (Sikorski *et al.*, 2013a), pigs, foxes (Bodewes *et al.*, 2013), bats (Castrignano *et al.*, 2013; Ge *et al.*, 2011; Ge *et al.*, 2012; Li *et al.*, 2010b), rodents (Phan *et al.*, 2011), mustelids (Smits *et al.*, 2013; van den Brand *et al.*, 2012) and birds (Phan *et al.*, 2013; Reuter *et al.*, 2014; Sikorski *et al.*, 2013b).

Several of these novel ssDNA viruses share similarities to *Sclerotinia sclerotiorum* hypovirulence-associated DNA virus 1 (SsHADV-1), a virus previously found in benthic river sediments (Kraberger *et al.*, 2013) and has been isolated from the fungus, *Sclerotinia sclerotiorum* (Yu *et al.*, 2010). The Rep of SsHADV-1 contains geminivirus-like Rep motifs (Dayaram *et al.*, 2012; Nash *et al.*, 2011). Viral genomes related to SsHADV-1 have also been recovered from animal faecal samples (Sikorski *et al.*, 2013b; van den Brand *et al.*, 2012), dragonflies and damselflies, mosquitoes (Ng *et al.*, 2011), bovine and human serum (Lamberto *et al.*, 2014) and plant material (Dayaram *et al.*, 2012; Du *et al.*, 2014) and, as a result, a new genus known as gemycircularviruses has been proposed for this group of Gemini-like, possibly fungal-infecting, viruses. Novel ssDNA viruses which have Reps that share similarity to circo- and cyclovirus proteins have also been identified in faecal samples (Ge *et al.*, 2011; Li *et al.*, 2010a; Li *et al.*, 2010b).

The genomes of monopartite eukaryote-infecting circular ssDNA viruses (such as those in the families *Circoviridae* and *Geminiviridae*) typically have at least two open reading frames (ORF): one encoding a Rep and the other encoding a coat protein (CP). Multipartite ssDNA virus genomes, such as those found in members of the *Nanoviridae*, are comprised of up to

eight individual genome components, with each component encoding a single protein; of those components at least one encodes a Rep and another component encodes a CP. The CPs of viruses are often highly diverse because they are involved in interaction with host cell surface receptors. In many instances the putative Rep encoded by divergent environmental circular Rep encoding single-stranded (CRESS) DNA viruses is the only protein that has any detectable homology to other known ssDNA virus Rep proteins. In fact this similarity is the strongest evidence that these environmental ssDNA molecules are virus genomes. Signature motifs within the Reps of these molecules are characteristic of Rep encoding ssDNA viruses (Rosario *et al.*, 2012b).

Sampling sewage provides a convenient and non-invasive approach to studying viral diversity within human impacted environments. As mentioned previously, aeration of treated sewage occurs in oxidation ponds where clarified sewage is circulated for ~two weeks before being discharged into the ocean or rivers. This open-air stage allows algal growth and UV exposure, both of which reduce coliform populations (Abdel-Raouf *et al.*, 2012; Sinton *et al.*, 2002). Despite a number of viral studies on raw sewage, only one study (Rosario *et al.*, 2009b) has looked at viral diversity in treated sewage prior to discharge. To address this lack of knowledge we used a viral metagenomic sequence-informed approach to determine the diversity of CRESS DNA viruses within treated sewage oxidation ponds. We recovered and characterised 50 novel ssDNA virus genomes and 11 sub-genome-length circular DNA molecules that may be either defective genomes or individual genome components of viruses belonging to divergent groups of multipartite ssDNA viruses. This highlights the rich diversity of the CRESS DNA viruses associated with sewage oxidation ponds.

8.3 Materials and methods

8.3.1 Sample collection and viral DNA isolation

A sewage oxidation pond (final ‘open air’ stage of treatment) water sample was collected at the Christchurch Wastewater Treatment Plant, Christchurch, New Zealand on September, 2012. 50ml of the sample was successively passed through 0.45µm and 0.2µm syringe filters (Sartorius Stedim Biotech, Germany). The filtrate was precipitated using 15% PEG at 4°C overnight and pelleted by centrifugation at 10,000 x g for 10min. The resulting pellet was resuspended in 1ml of SM buffer [0.1 M NaCl, 50 mM Tris-HCl (pH 7.4)]. Nucleic acid was extracted from 200µl of this re-suspension using the High Pure Viral Nucleic Acid Kit (Roche Diagnostics, USA). The isolated viral nucleic acids were enriched using TempliPhi™ (GE Healthcare, USA).

8.3.2 Next-generation sequencing-informed recovery of complete viral genomes

Enriched viral DNA was sequenced at the Beijing Genomics Institute (Hong Kong) using an Illumina HiSeq 2000 sequencer. The paired end reads were assembled using ABySS 1.3.5 (Simpson *et al.*, 2009) with a k-mer setting of 64. Contigs >1000nts were analysed by BLASTx (Altschul *et al.*, 1990) for detectable homology to known viral proteins.

Abutting primers were designed (Table 1) to recover complete circular genomes for contigs found to have credible viral hits BLAST E-scores $<10^{-7}$) as determined using BLASTx (Altschul *et al.*, 1990). The circular genomes were recovered using polymerase chain reaction with KAPA HiFi Hotstart DNA polymerase (Kapa Biosystems, USA) with the specific abutting primers using the following thermocycler program: 94°C for 3 min, 25 cycles of 98 °C (20sec), 55 °C (30sec), 72 °C (3min) and a final extension of 72°C for 3 min. The PCR amplicons were gel purified and ligated into pJET1.2 plasmid (Thermo Fisher Scientific, USA) and sequenced at Macrogen Inc. (Korea) by Sanger sequencing using primer walking.

Sanger sequencing reads were assembled using DNA Baser (Heracle BioSoft S.R.L. Romania). Putative CP and Rep ORFs were identified and preliminary genome analysis was carried using BLASTx (Altschul *et al.*, 1990). Pairwise similarity comparisons (1 – p-distance, with pairwise deletion of gaps) of Rep amino acid sequences predicted to be

expressed by circular CRESS DNA viruses obtained in this study along with those available in GenBank (as of 14th Sept 2014), were determined using SDT v1.2 (Muhire *et al.*, 2014).

Table 8.1: Details of primer sequences used to recover complete CRESS DNA viral genomes and circular DNA molecules in this study.

Sewage-associated circular DNA viruses			
Acronym and isolate details	GenBank accession #	Forward primer	Reverse primer
SaCV-1 [NZ-BS3349-2012]	KJ547620	GTGACAAGTACAAACGAAAACG	ATACCACCAGCCAGTCTC
SaCV-2 [NZ-BS4000-2012]	KJ547626	TCACAGCGAGGGTAAGTTAAG	CTTAGACTTGCAACTACTTCTTG
SaCV-3 [NZ-BS3854-2012]	KJ547627	ACGACTTCGAAACGGTCTC	CGTTTCTTGCGGTATTTGTCAGT
SaCV-4 [NZ-BS3799-2012]	KJ547628	GGCGAACTTTAAGGATGACTG	AAATGGTTGAAGTACATGTGCGG
SaCV-5 [NZ-BS3901a-2012]	KJ547629	GGTGGCTGTTTCGTAAGGATA	CACTGAAATGGTATGATTGGAC
SaCV-6 [NZ-BS4017-2012]	KJ547630	GAGCTTGTTTGACCTTCTTCTC	TCGAAGCGGAGGACATTGAA
SaCV-7 [NZ-BS3976-2012]	KJ547631	TTCCTTCCAAACGTAGTCACTG	GATACTCGAGTCTCCGGAA
SaCV-8 [NZ-BS4075-2012]	KJ547632	TCGGTGGTCTTGATAGCT	CTGGGACGACATCTGGAAAT
SaCV-9 [NZ-BS3681-2012]	KJ547633	GCAGTCGCAGAAAGAAACG	TACCCGTGACTCCCAGATT
SaCV-10 [NZ-BS3946-2012]	KJ547621	ACCTACGTACTCCTCAGC	TACGGAGTTATCGAGCAGTTG
SaCV-11 [NZ-BS3997-2012]	KJ547622	CACTATTCGTACCTACACTTGG	CTAGTACGGATCCTGTTGTATTG
SaCV-12 [NZ-BS3888-2012]	KJ547623	CCATGCATCTGGCTCAACAA	CTTTCCCCATCTGACTGTTT
SaCV-13 [NZ-BS4044-2012]	KJ547624	TGAGATTACGAGCAATCTTCAC	CAAGGATGGAAGCAAGCATATG
SaCV-14 [NZ-BS4064-2012]	KJ547625	TTAGCCTCCCAGAGAGAGA	GCTCTCCTGGGCTGTTGT
SaCV-15 [NZ-BS3557-2012]	KM821750	TCGTTGGGTCTCTACGGAGTTTG	CCCGAACCGTAAACTGATAGTC
SaCV-16 [NZ-BS3759-2012]	KM821751	GACGCCATCAGAGATGCAGCT	CCAATCTCGAGAGTCATTGCGG
SaCV-17 [NZ-BS4236-2012]	KM821752	CAACCCAGGAACCTATTACTTCTCC	GATTCTGCGAATCTTCTAACACCTCC
SaCV-18 [NZ-BS3994-2012]	KM821753	CAGGGTAACTATCAGGTTACAAAGAG	CTTGCCACCAATAAAGTTGAAACAGTG
SaCV-19 [NZ-BS4128-2012]	KM821754	TCGCAGTCGATAATATGGCGCC	TTCGTTTGAGGTTCCCGTACAATCTAC
SaCV-20 [NZ-BS3900-2012]	KM821755	AGCTCATCCATTGCAGCAGCAC	TATGGGTCATACCATTGAGTGATACG
SaCV-21 [NZ-BS4169-2012]	KM821756	CACTGGCGAGTGGTTCTATGG	ACGTCCTGCTGATCCTCAGG
SaCV-22 [NZ-BS4155-2012]	KM821757	GAGCTCTCCACTCGAGAGTTC	GCGATGAAAAGACTGAAGTCGTGG
SaCV-23 [NZ-BS4025-2012]	KM821758	GAAGTGACTTGCTTAGAGTCGCTG	TCTGTCCGGCTCCTGTACAG
SaCV-24 [NZ-BS4091-2012]	KM821759	GTAGAACGATAGCCCAGTCGG	AAGCTGCCAGAGAAAGAAGACTTGG
SaCV-25 [NZ-BS4281-2012]	KM821760	GGCCATGTACCGTTCGATCTG	CTGGTCGAGGACGAAGTACC
SaCV-26 [NZ-BS4339-2012]	KM821761	GAAGTGGATGACCTTGTTCTCTGG	ATCCAGAGTCTTGCCAGCTTTGATC
SaCV-27 [NZ-BS4103-2012]	KM821762	CAGCGTGAATTCGAAACCAC	ATAGCTCTGAGAGTCCGATAACTGTG
SaCV-28 [NZ-BS4064a-2012]	KM821763	CAAGCTGGAGGAGTTCATTGGATG	CGCTTCTGGAGCATGCAGTAAC
SaCV-29 [NZ-BS4325-2012]	KM821764	GATGAAAGCGCGAACTGACTATCC	GTTTTGCCCGCACCAGATGCT
SaCV-30 [NZ-BS4120-2012]	KM821765	CGTGGATTGGCGAGAAAAGCTC	TAAGCTTGCATTGCTTTCCGTGGAG

Table 8.1 continued

Acronym and isolate details	GenBank accession #	Forward primer	Reverse primer
SaCV-31 [NZ-BS4358-2012]	KM821766	CATGGAACAACCCCAACGTTACTG	TAACACACCAACGCATGACTTTAGGC
SaCV-32 [NZ-BS4194-2012]	KM821767	GGCTGGATCGGTACCCAGTAAT	ATCGTAAAAAGTGCAGACATCGATTC
SaCV-33 [NZ-BS4147-2012]	KM821768	GGCAAAAGTCATCGTGCAAGATCTG	AACACCTGGAGGCCCGTAATAC
SaCV-34 [NZ-BS4221-2012]	KM821769	TTACGTCGACAAAAGACGCTGACAAG	CAAAGCAACTCTTCTTGCCATTTTCG
SaCV-35 [NZ-BS4050-2012]	KM821770	GCCATTGCCTGCTACGCCTA	GTGAGCGATGCCTTTGCAGGTTT
SaCV-36 [NZ-BS3974-2012]	KM821748	CGCAACGACAGCATCGTCAAG	GAGGACCCACTCTGTAGGCT
SaCV-37 [NZ-BS2945-2012]	KM821749	ACTTTATGCTCCTCGGGTGCAG	GGTCTTGGAACCATTGGAGGATAAC
Sewage-associated gemycircularvirus			
SaGmV-1 [NZ-BS3970-2012]	KM821747	GAATGGCTATTACAGTCTGGTATCGG	ATCTCTTCCATCAACATCTCCTCCG
SaGmV-2 [NZ-BS3911-2012]	KJ547642	AATCTCCGTTACCCCTCTTTC	GTGAAAACTAAAGTCCGAAGG
SaGmV-3 [NZ-BS4149-2012]	KJ547643	ACAGAAGTGCCCTTGGTG	GTTCAATCACCTCACTCCG
SaGmV-4 [NZ-BS3913-2012]	KJ547634	AGGATGGAGGAGTTCATTAC	GGTGACGTTCTCTTCCCA
SaGmV-5 [NZ-BS3963-2012]	KJ547635	TACCACCCGAACATTCAACCA	GCCTTCAACATCAAAGCGTC
SaGmV-6 [NZ-BS4014-2012]	KJ547636	GGTTACGACTACGCATGCAA	CTTCTCCGGAGTGCCATAA
SaGmV-7a [NZ-BS3939-2012]	KJ547637	TCCATCCGCGTGATCTTCT	GGAGTTCATCTGCACGTG
SaGmV-7b [NZ-BS3972-2012]	KJ547640	ACACAGATCGACGATGGTAC	AAAATATCGCGCGAGATTCGG
SaGmV-8 [NZ-BS3917-2012]	KJ547638	AGTACTGTGACTGGAAGTTCG	CATGTTCTTTACCCAACTTCAGA
SaGmV-9 [NZ-BS3970-2012]	KJ547639	GCCATTGCTTCCTCCGCT	GGAACCTCGTGATAAGTGGG
SaGmV-10a [NZ-BS3980-2012]	KJ547644	GTCTTTCGAGTCCCCCA	CTATCGGAGGCAGAGTG
SaGmV-10b [NZ-BS3849-2012]	KJ547645	ACTCTGCTTCCTGAAGGTC	GGATGCAAATGGTGGCGT
SaGmV-11 [NZ-BS4117-2012]	KJ547641	CCTTGATTGCATAGTCGTATCC	ATGGTGATGTTGTTTGC GGAG
Sewage-associated circular DNA molecules			
SaCM-1 [NZ-BS4111-2012]	KJ547618	TACTGTACCAAAGAAGAAGGTCGC	CTTTTCATTTTGTTGCTTGGTACCC
SaCM-2 [NZ-BS3901b-2012]	KJ547617	GGTGGCTGTTCTGAAGGATA	CACTGAAATGGTATGATTGGAC
SaCM-3 [NZ-BS2940-2012]	KJ547619	ATCTTCTTGCCGACCCGTT	TATGCTGAAGAGTCTCCAAGC
SaCM-4 [NZ-BS2920-2012]	KM877826	CCTTTGCCGATTACGCAAACACTAG	CTGCTATGACCTTGCCATCGTTC
SaCM-5 [NZ-BS3056-2012]	KM877827	CTGCTACCTGGATGTCATGTAGAG	AGCTATAACCTGGGCTAAGGACTTC
SaCM-6 [NZ-BS3713-2012]	KM877828	ACAGCTGCCACTGCTGTTTTCC	CTGAGCCGCAGGATCAACAGT
SaCM-7 [NZ-BS3510-2012]	KM877829	GATATTCTGCAGCCCCAACTTC	TCCTATCTATCGCGTAGAATATAGGAGTC
SaCM-8 [NZ-BS3610-2012]	KM877830	GATCCTTCGGGAGAAGCCGAA	TGCCGTACCGCAGTCCAGAAT
SaCM-9 [NZ-BS3553-2012]	KM877831	CGACTATCATTGGTACACTACCTAACC	GCTGAGTATCAGGTACACGAGTG
SaCM-10 [NZ-BS3301-2012]	KM877832	GATCACCTGGATCTATGATCACAATAGC	TCGATCTTCTGATCTTGATCATGTCG
SaCM-11 [NZ-BS3394-2012]	KM877833	GACATCGCACACGGGAACAACAA	GTCGACGAAGGAGTCTTGACAGAGAG

8.3.3 Phylogenetic analyses

The Rep amino acid sequences potentially encoded by the newly determined CRESS virus genomes together with those potentially encoded by other CRESS DNA viruses available in Genbank (as of 14th Sept 2014) were aligned using T-coffee (Notredame *et al.*, 2000) and the alignment was refined using MUSCLE (Edgar, 2004) as implemented in MEGA5 (Tamura *et al.*, 2011).

The refined alignment was used to infer a maximum likelihood (ML) phylogenetic tree using PhyML v3 (Guindon *et al.*, 2010) using the best fit amino acid substitution model, WAG+G (determined using ProtTest 3) (Darriba *et al.*, 2011), with an approximate likelihood-ratio test (aLRT) (Anisimova & Gascuel, 2006) for branch support. All branches with <80% aLRT support were collapsed using Mesquite v2.75 (<http://mesquiteproject.org>).

A specific Rep alignment was generated which contained all known gemycircularvirus-like Rep sequences and all other available Rep sequences that were closely related to this group. The Rep ML phylogenetic tree constructed using this dataset was inferred with PHYML using the LG+I+G amino acid substitution model (determined to be the best fit model with ProtTest 3). The Reps of nanoviruses were used as an out-group. Additionally, the Reps of DflaCV-1, DflaCV-2, RW-E, CB-B, RodSCV-M-44, BtCV-1, 12-LDMD, 18-LDMD, SI00898, OdasCV-12, SaCV-6, -7, -8, -16, -19, -24, -27 and -32 (all referred to as CRESS DNA clade-1) were aligned and a ML phylogenetic tree was constructed using the LG+I+G amino acid substitution model (determined to be the best fit by ProtTest 3). This tree was rooted with GOM03193 (JX904377) - a ssDNA viral sequence isolated from a seawater sample which is the most closely related outlier sequence to this group.

8.3.4 Recombination analysis of gemycircularviruses

Recombination analysis was performed on the aligned full genome dataset of the gemycircularviruses (as this group of CRESS DNA viruses are similar enough to enable a credible alignment) using RDP4 (Martin *et al.*, 2010) implementing the following methods: RDP (Martin & Rybicki, 2000), GENECONV (Padidam *et al.*, 1999), Bootscan (Martin *et al.*, 2005), Maxchi (Smith, 1992), Chimera (Posada & Crandall, 1998), Siscan (Gibbs *et al.*,

2000), and 3Seq (Boni *et al.*, 2007). Only events in which recombination was detected by three or more methods, each with an associated p-value of $<10^{-3}$ that were coupled with phylogenetic support for recombination having occurred, were considered credible.

8.4 Results and discussion

8.4.1 Recovery and characterisation of novel CRESS DNA viral genomes

8.4.1.1 Circular viral genomes recovered by PCR using abutting primers

In an initial study to investigate CRESS DNA viral diversity in treated sewage, we purified viral DNA from an oxidation pond sample, amplified sequenced this using an Illumina next-generation sequencing platform. The resulting 26,757,090 Illumina reads which were *de novo* assembled into 1833 contigs that were >1000 nts. A BLASTx comparison of these reads to the NCBI viral protein database showed that 53.3% (977 contigs) had significant BLASTx hits (e-value of 10^{-7}). Of those 95.7% (935 contigs) shared identity with viral protein. Further analysis of the 935 viral-like contigs showed that 39.04% (365 contigs) shared identity to CRESS DNA virus proteins, 30.7% (287 contigs) to *Microviridae* proteins and 23.42% (219 contigs) to *Inoviridae* proteins and 6.93% (64 contigs) to double stranded DNA viruses (*Phycodnaviridae*, *Mimiviridae*, *Irodoviridae*, *Corticoviridae*, *Siphoviridae*, *Podoviridae* and *Mucoviridae*). For the purpose of this study we focused on the 365 contigs with similarities to CRESS DNA viruses. Contigs were used to design back-to-back primers within a conserved region of the Rep or CP encoding ORF sequences. These primers were used in a PCR to recover circular ssDNA molecules representing what are apparently complete genomes of 50 novel ssDNA viruses (Table 1). These sequences are hereafter simply referred to as “complete genomes” although it is recognised that these could possibly be genome components of multipartite ssDNA viruses such as those found in the genus *Begomovirus* of the family *Geminiviridae* and in all genera of the family *Nanoviridae*. Eleven additional circular replicons were recovered that are either individual genome components of multipartite ssDNA viruses, or defective sub-genome-length molecules that are hereafter simply referred to as “subgenomic DNA molecules”.

All the PCR products were cloned and Sanger sequenced (a list of the primers used to recover each circular DNA molecule is provided in Table 1). Thirty-seven of the complete genomes

that range in size from 1605 nts to 4202 nts likely belong to completely novel CRESS DNA virus genera as they share only very low levels of sequence similarity to any other known CRESS DNA viruses. We have therefore tentatively named these sequences Sewage-associated circular DNA virus (SaCV) 1 through 37. All of these CRESS DNA viruses encode at least two major ORFs, one of which encodes a Rep (Fig. 8.1). SaCV-1, SaCV-2, SaCV-3, SaCV-24, SaCV-32 SaCV-35, SaCV-36, SaCV-37 likely express Reps from spliced transcripts (i.e. their *rep* genes likely have introns analogous/homologous to those found in some geminiviruses and gemycircularviruses). Reps derived from a spliced *rep* are a feature seen in some members of the *Geminiviridae* family, for example in the genus *Mastrevirus* the *rep* is expressed from a spliced ORF C1 and C2 and *repA* is expressed from ORF C1 alone (Dekker *et al.*, 1991; Wright *et al.*, 1997). Splicing event of the Rep occurs with the removal of an intron with the acceptor and donor sites of GT and AG, respectively. In geminivirus the *rep* and *repA* are both essential for replication (Liu *et al.*, 1998), the *repA* also plays a role in gene expression (Collin *et al.*, 1996; Liu *et al.*, 1999). Large ORFs that potentially encode Rep and CP are identifiable in all of the SaCV sequences other than SaCV-8 and -31 (See Table 2 for the top BLASTx result for each of the SaCVs putative Rep and CP ORFs). In these exceptional viruses two large ORFs other than the Rep were identified, neither of which shared any similarities to known CRESS DNA virus CPs. Similarly, two viruses with two unknown ORFs in addition to a ORF which encodes a Rep, have also been identified from a fresh water pond in the McMurdo Ice Shelf (Antarctica) (Zawar-Reza *et al.*, 2014). Of the 37 SaCV genomes 17 have putative Rep and CP ORFs which are bidirectionally transcribed (SaCV-31 has two unknown ORF, both transcribed on the virion sense strand), whereas 19 are unidirectionally transcribed and one SaCV-8 is unknown due to two possible CP ORFs transcribed both on the virion and complementary strands. Similar genome architecture has been documented in other CRESS DNA viruses (Dayaram *et al.*, 2014; Labonté & Suttle, 2013; Rosario *et al.*, 2009a; Zawar-Reza *et al.*, 2014). These 37 complete genomes are so divergent from all other known CRESS DNA viruses, as is the case for most environmental CRESS DNA viruses, we were therefore only able to analyse their phylogenetic relationships to the CRESS DNA viruses based on their predicted Rep amino acid sequences (Fig. 8.2; additional Table 8.1 for list of acronyms and corresponding accession numbers in phylogenetic tree).

The remaining 13 complete genomes that were recovered here are closely related to the gemycircularviruses (Fig. 8.2 and 8.3), all have two large bidirectional ORFs (putatively encoding a Rep and CP) and ranging in size from 2130 nts to 2277 nts (Fig. 8.1). As is the case with most other described gemycircularviruses, all but one (SaGmV-6) of these complete genomes likely express Rep from a spliced transcript (Fig. 8.1).

Genome-wide pairwise nucleotide comparisons of the gemycircularviruses which include sequences of 13 genomes recovered in this study together with 37 related sequences currently available in GenBank revealed a similar distribution of pairwise identities to that determined by Sikorski et al (2013b). Specifically, the distribution of pairwise identity scores display discrete clusters between 55% – 67%, 70% – 76% and 78% – 81%. Based on this distribution (Fig. 8.4), coupled with phylogenetic evidence, (Fig. 8.3A) we have tentatively classified the new viral genome sequences into separate species based on a genome-wide nucleotide sequence identity threshold of <78%. Hence, 13 of the gemycircularviruses recovered in this study were classified into 11 new species. Following from the simplified nomenclature proposed by Sikorski et al (2013b) we tentatively name these viruses sewage-associated gemycircularvirus (SaGmV) -1 through -11. The two genomes assigned to the SaGmV-7 “species” (genomes BS3939 and BS3972) share 79.2% genome-wide nucleotide identity, and the two assigned to the SaGmV-10 species (genomes BS3980 and BS3849) share 81.1% identity. We have tentatively named these SaGmV-7 and SaGmV-10 sequences, SaGmV-7a / 7b, and SaGmV-10a / 10b, respectively.

8.4.1.2 Subgenomic DNA molecules

The 11 subgenomic DNA molecules that were recovered (Fig. 8.1) which range in size from 896 nts to 1294 nts and may represent either defective genomes (with large deletions) or small genome components of CRESS DNA viruses with multipartite genomes such as those found in the family *Nanoviridae*. We identified a single large ORF in each of these DNA molecules and have tentatively named them Sewage-associated circular DNA molecules (SaCM) -1 through to -11. Eight of these subgenomic molecules have a single ORF which shares similarity to Reps and three have an ORF that shares similarities to CPs of CRESS DNA viruses. A putative virion strand origin of replication stem-loop structures such as those found in most other known CRESS DNA viruses were identifiable within 7 of the 11 subgenome molecules. Given both the absence of easily identifiable virion strand origins of

replication in all subgenomes and the fact that all 11 of these sequences are most closely related to CRESS DNA viruses with genomes that are twice as large and likely express two proteins, it is likely that these molecules are non-viable defective versions of much larger genomes rather than the complete components of multipartite genomes. The only exception to this is SaCM-4 which shares 53% identity with *Faba bean necrotic yellows virus* DNA R (KC979000) (Table 2). Geminivirus subgenomes have been recorded as a natural occurrence during geminivirus replication (Hadfield *et al.*, 2012; Jeske *et al.*, 2001), it is therefore not surprising that other ssDNA viruses may also produce subgenomic molecules during replication.

Table 8.2: Top BLASTx identities of the sewage-associated viruses and molecules Rep and CP amino acid sequences determined in this study with those encoded by other complete CRESS DNA virus genomes.

Sewage-associated circular DNA viruses									
Isolate name	GenBank accession #	Replication-associated protein	GenBank accession #	E-Value	Identity	Coat protein	GenBank accession #	E-Value	Identity
SaCV-1	KJ547620	Baminiavirus	JQ898331	9×10^{-30}	32%	McMurdo Ice Shelf pond-associated circular DNA virus-3	KJ547648	2×10^{-38}	31%
SaCV-2	KJ547626	Gemycircularvirus-10	KF371632	6×10^{-28}	50%	no significant viral hit			
SaCV-3	KJ547627	Baminiavirus	JQ898331	1×10^{-39}	35%	no significant viral hit			
SaCV-4	KJ547628	Gemycircularvirus-11	KF371631	5×10^{-55}	38%	no significant viral hit			
SaCV-5	KJ547629	Chimp162	GQ404883	5×10^{-07}	45%	no significant viral hit			
SaCV-6	KJ547630	Circoviridae 18 LDMD	KF133825	5×10^{-137}	65%	no significant viral hit			
SaCV-7	KJ547631	Dragonfly larvae associated circular virus-2	KF738874	2×10^{-141}	67%	no significant viral hit			
SaCV-8	KJ547632	Dragonfly larvae associated circular virus-2	KF738874	9×10^{-97}	58%	no significant viral hit			
SaCV-9	KJ547633	Pig stool associated circular ssDNA virus	JX305992	8×10^{-136}	78%	Turkey stool associated circular ssDNA virus	KF880727	2×10^{-116}	55%
SaCV-10	KJ547621	Pea yellow stunt virus	KC979054	3×10^{-30}	55%	no significant viral hit			
SaCV-11	KJ547622	Circoviridae 2 LDMD	KF133808	3×10^{-18}	43%	Circoviridae 4 LDMD-2013			
SaCV-12	KJ547623	Bat circovirus ZS/China/2011	JF938079	9×10^{-52}	40%	no significant viral hit			
SaCV-13	KJ547624	HCB18.215 virus	LK931483	5×10^{-07}	60%	Circoviridae 2 LDMD-2013	KF133808	2×10^{-20}	42%
SaCV-14	KJ547625	Tobacco yellow dwarf virus-A	JN989443	4×10^{-15}	27%	Circoviridae 21 LDMD-2013	KF133828	4×10^{-13}	32%
SaCV-15	KM821750	Bat circovirus ZS/Yunnan-China/2009	JN377572	5×10^{-08}	35%	uncultured marine virus	JX904404	1×10^{-24}	35%
SaCV-16	KM821751	Dragonfly larvae associated circular virus-2	KF738874	4×10^{-109}	61%	McMurdo Ice Shelf pond-associated circular DNA virus-7	KJ547652	4×10^{-10}	27%

Table 8.2 continued

Isolate name	GenBank accession #	Replication-associated protein	GenBank accession #	E-Value	Identity	Coat protein	GenBank accession #	E-Value	Identity
SaCV-17	KM821752	Circoviridae 19 LDMD	KF133826	3x10 ⁻⁴¹	37%	no significant viral hit			
SaCV-18	KM821753	Rhynchosia golden mosaic virus	AF239671	2x10 ⁻¹⁴	26%	Circoviridae 4 LDMD-2013	KF133811	6x10 ⁻¹⁷	30%
SaCV-18	KM821754	Bat circovirus ZS/China/2011	JF938078	2x10 ⁻¹⁰⁶	57%	Nepavirus	JQ898333	3x10 ⁻¹⁶	28%
SaCV-20	KM821755	Circoviridae 19 LDMD-2013	KF133826	6x10 ⁻⁶⁵	44%	no significant viral hit			
SaCV-21	KM821756	Dragonfly larvae associated circular virus-3	KF738876	7x10 ⁻⁷⁶	47%	no significant viral hit			
SaCV-22	KM821757	Canine circovirus	KC241983	2x10 ⁻⁵⁹	43%	no significant viral hit			
SaCV-23	KM821758	Dragonfly cyclovirus1	KC512918	2x10 ⁻⁷⁵	45%	no significant viral hit			
SaCV-24	KM821759	Bat circovirus ZS/China/2011	JF938078	4x10 ⁻⁸⁶	49%	no significant viral hit			
SaCV-25	KM821760	Farfantepenaeus duorarum circovirus	KC441518	4x10 ⁻²⁸	35%	Circoviridae 21 LDMD-2013	KF133828	3x10 ⁻⁰⁶	29%
SaCV-26	KM821761	Gemycircularvirus-9	KF371633	8x10 ⁻¹⁰	25%	Dragonfly larvae associated circular virus-3	KJ547622	1x10 ⁻⁰⁶	27%
SaCV-27	KM821762	Dragonfly larvae associated circular virus-2	KF738874	5x10 ⁻¹²⁵	65%	Circoviridae 18 LDMD-2013	KF133825	3x10 ⁻¹⁶	28%
SaCV-28	KM821763	Citrus chlorotic dwarf associated virus	KJ547625	4x10 ⁻¹⁴	30%	Mosquito VEM virus SDRBAJ	HQ335087	2x10 ⁻²⁴	34%
SaCV-29	KM821764	Diporeia sp. associated circular virus	KC248416	7x10 ⁻⁶⁰	43%	Circoviridae 21 LDMD-2013	KF133828	1x10 ⁻⁰⁹	28%
SaCV-30	KM821765	Chickpea chlorosis Australia virus	KC172693	2x10 ⁻¹⁰	27%	Dragonfly larvae associated circular virus-3	KF738876	2x10 ⁻⁰⁷	29%
SaCV-31	KM821766	Circovirus-like genome RW-A	FJ959077	4x10 ⁻²⁶	30%	no significant viral hit			
SaCV-32	KM821767	Dragonfly larvae associated circular virus-2	KF738874	2x10 ⁻¹⁰⁷	58%	Nepavirus	JQ898333	6x10 ⁻¹⁴	24%

Table 8.2 continued

Isolate name	GenBank accession #	Replication-associated protein	GenBank accession #	E-Value	Identity	Coat protein	GenBank accession #	E-Value	Identity
SaCV-33	KM821768	Dragonfly larvae associated circular virus-3	KF738876	1×10^{-53}	43%	Nepavirus	JQ898334	3×10^{-19}	28%
SaCV-34	KM821769	Cyanoramphus nest associated circular X DNA virus	JX908739	3×10^{-36}	32%	no significant viral hit			
SaCV-35	KM821770	Cyclovirus PKbeef23/PAK/2009	HQ738634	1×10^{-21}	29%	Circoviridae 3 LDMD-2013	KF133810	1×10^{-09}	30%
SaCV-36	KM821748	Ancient caribou feces associated virus	KJ938716	5×10^{-146}	60%	Ancient caribou feces associated virus	KJ938716	1×10^{-71}	50%
SaCV-37	KM821749	Dragonfly-associated circular virus-1	JX185430	3×10^{-52}	35%	no significant viral hit			
Sewage-associated gemycircularviruses									
SaGmV-1	KM821747	Cassava associated circular DNA virus	JQ412057	1×10^{-142}	63%	Gemycircularvirus-5	KF371637	6×10^{-37}	35%
SaGmV-2	KJ547642	Dragonfly-associated circular virus-1	JX185430	6×10^{-122}	59%	MSSI2.225 virus	LK931485	5×10^{-43}	41%
SaGmV-3	KJ547643	Gemycircularvirus-8	KF371634	0	67%	Gemycircularvirus 8	KF371634	6×10^{-141}	62%
SaGmV-4	KJ547634	Caribou feces-associated gemycircularvirus	KJ938717	2×10^{-94}	63%	Caribou feces-associated gemycircularvirus	KJ938717	1×10^{-89}	50%
SaGmV-5	KJ547635	Meles meles fecal virus	JN704610	5×10^{-128}	73%	Meles meles fecal virus	JN704610	8×10^{-129}	57%
SaGmV-6	KJ547636	Sclerotinia sclerotiorum hypovirulence associated DNA virus 1	GQ365709	0	71%	Gemycircularvirus-10	KF371632	8×10^{-57}	44%
SaGmV-7a	KJ547637	Dragonfly-associated circular virus 2	JX185429	9×10^{-161}	65%	Sclerotinia sclerotiorum hypovirulence associated DNA virus 1	GQ365709	5×10^{-18}	50%
SaGmV-7b	KJ547640	Cassava associated circular DNA virus	JQ412056	6×10^{-149}	62%	Sclerotinia sclerotiorum hypovirulence associated DNA virus 1	GQ365709	2×10^{-19}	49%

Table 8.2 continued

Isolate name	GenBank accession #	Replication-associated protein	GenBank accession #	E-Value	Identity	Coat protein	GenBank accession #	E-Value	Identity
SaGmV-8	KJ547638	Cassava associated circular DNA virus	JQ412057	8x10 ⁻¹⁵⁹	62%	Cassava associated circular DNA virus	JQ412057	4x10 ⁻³⁶	40%
SaGmV-9	KJ547639	Dragonfly-associated circular virus-2	JX185429	2x10 ⁻¹⁵⁴	62%	Gemycircularvirus-4	KF371638	4x10 ⁻¹⁴	41%
SaGmV-10a	KJ547644	Gemycircularvirus-9	KF371633	7x10 ⁻¹⁶⁹	65%	MSSI2.225 virus	LK931484	2x10 ⁻¹⁵⁶	80%
SaGmV-10b	KJ547645	MSSI2.225 virus	LK931485	9x10 ⁻¹³⁰	94%	MSSI2.225 virus	LK931485	2x10 ⁻¹⁷⁴	88%
SaGmV-11	KJ547641	Hypericum japonicum associated circular DNA virus	KF413620	2x10 ⁻¹⁴⁴	61%	Gemycircularvirus-2	KF371640	1x10 ⁻⁵⁵	45%
Sewage-associated circular DNA molecules									
SaCM-1	KJ547618	Beak and feather disease virus	JX221020	4x10 ⁻¹²	47%	No putative Cp			
SaCM-2	KJ547617	Diporeia sp. associated circular virus	KC248418	3x10 ⁻¹⁸	30%	No putative Cp			
SaCM-3	KJ547619	Circoviridae 4 LDMD	KF133811	5x10 ⁻¹⁴	31%	No putative Cp			
SaCM-4	KM877826	Faba bean necrotic yellows virus	KC979000	1x10 ⁻²⁶	53%	No putative Cp			
SaCM-5	KM877827	Dragonfly larvae associated circular virus-3	KF738876	7x10 ⁻⁶³	49%	No putative Cp			
SaCM-6	KM877828	No putative Rep				Circovirus-like genome RW-C	FJ959079	3x10 ⁻⁰⁴	62%
SaCM-7	KM877829	No putative Rep				Circovirus-like genome RW-C	FJ959079	7x10 ⁻¹⁹	34%
SaCM-8	KM877830	Bat circovirus ZS/Yunnan-China/2009	JN377572	8x10 ⁻⁰⁷	36%	No putative Cp			
SaCM-9	KM877831	Meles meles circovirus-like virus	JQ085285	1x10 ⁻²⁶	32%	No putative Cp			
SaCM-10	KM877832	No putative Rep	-			Diporeia sp. associated circular virus	KC248418	4x10 ⁻⁰⁹	26%
SaCM-11	KM877833	Anguilla anguilla circovirus	KC469701	3x10 ⁻²⁹	31%	No putative Cp			

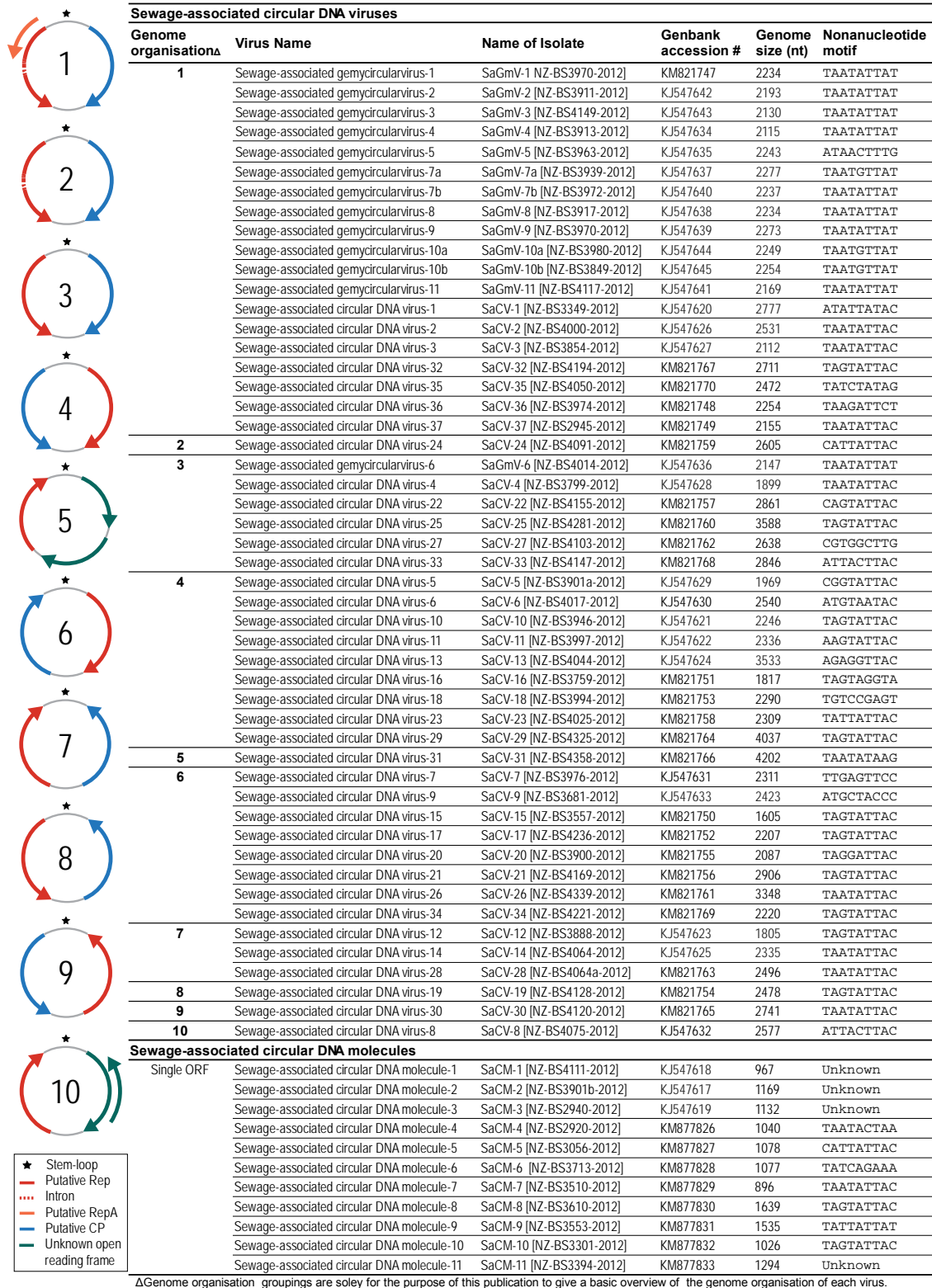


Figure 8.1: Summary of genome organisation and isolate information for sewage-associated viruses and molecules. Genome cartoons show overview of genome organisation but do not represent the exact positioning and relative size of each gene in the individual sewage associated viruses.

8.4.2 Phylogenetic and sequence analyses of novel sewage-associated circular viruses

The only recognisably homologous features in common between the CRESS DNA viral genomes recovered in this study are their probable Reps and virion-strand origins of replication. An alignment containing predicted Rep amino acid sequences from both the probable viral genomes recovered in this study and those of other representative CRESS DNA viruses available in GenBank was used to construct a ML phylogenetic tree. It is important to note that the breadth of diversity in this large dataset makes it difficult to accurately align the Rep sequences and hence the ML phylogenetic tree presented in Figure 8.2 is simply meant as a plausible indicator of the genetic relatedness between the represented Rep sequences of CRESS DNA viruses.

The Reps of the SaCVs are highly diverse, with SaCV-1, -2, -3, -4, -36 and -37 being most closely related to geminiviruses, the gemycircularviruses (Rosario *et al.*, 2012a; Sikorski *et al.*, 2013b; Yu *et al.*, 2010), two other sewage-associated viruses, (Baminivirus and Niminivirus) (Ng *et al.*, 2012), an Ancient caribou-associated virus (anCFV) (Ng *et al.*, 2014) and an Odonata-associated circular DNA virus (OdasCV-6) (Dayaram *et al.*, In review) (Fig. 8.2 and Fig. 8.3). SaCV-1 encodes a putative CP which shares 29.31% amino acid identity with the presumed CP of Niminivirus. The SaCV-9 Rep appears to be clustering with a group of divergent CRESS DNA viruses isolated from pig faeces, sharing ~79% Rep aa identity with pig stool associated circular DNA viruses from Europe (JX305991- JX305998, unpublished and JQ023166, (Sachsenröder *et al.*, 2012). The Rep of SaCV-10 is most similar to that of nanoviruses, sharing 42.65% pairwise identity with *Milk vetch dwarf virus* (AB000920). Interestingly, the genome of SaCV-13 shares some similarity to the DNA-RNA hybrid viruses and viral contigs (Diemer & Stedman, 2012; McDaniel *et al.*, 2013; Roux *et al.*, 2013) recently described. The Rep of this virus is most closely related to that of ssDNA viruses whereas the CP is most similar to that of Tombusviruses, the oomycete-infecting ssRNA viruses, *Sclerophthora macrospora* virus A (Yokoi *et al.*, 1999) and *Plasmopara halstedii* virus A (Heller-Dohmen *et al.*, 2011), and the CP of previously described DNA-RNA viruses and viral-like contigs (Diemer & Stedman, 2012; McDaniel *et al.*, 2013; Roux *et al.*, 2013). Similar putative DNA-RNA hybrid viruses have been identified associated with dragonflies (Rosario *et al.*, 2012a) and in ocean water (McDaniel *et al.*, 2013) however they vary in that the CP of these viruses has a domain which is significantly similar to that in the

ssRNA plant-infecting virus, *Tobacco necrosis satellite virus* (Henriksson *et al.*, 1981). Five SaCVs, SaCV-11, -14, -18 and -29 all have CPs which also share significant similarities to STNV. The presence of these putative DNA/RNA hybrid viruses in several aquatic environments and from a dragonfly indicates that these viruses may be common in nature.

The Reps of SaCV-6, -7, -8 -16, -19, -24, -27 and -32 are all quite closely related to those of DflaCV-1 (KF738873), DflaCV-2 (KF738874), RW-E (FJ959081), CB-B (FJ959083), RodSCV-M-44 (JF755408), YN-BTCV-1 (JF938078), 12-LDMD (KF133819), 18-LDMD (KF133825), OdasCV-12 (KM598395) and SI00898 (JX904478; Fig. 8.2 and Fig 8.5). A more comprehensive analysis of this clade, which we have name CRESS DNA viruses Clade 1 for the purposes of this study, shows that the SaCVs in this clade share between 40% and 73% Rep aa identity with the other members of this clade and each other (Table 2). It is important to note that despite these genomes having closely related Rep sequences their genome organisations vary and are still highly diverse which is evident in the ML phylogenetic tree (Fig. 8.5). Those which have a unidirectional genome organisation are OdasCV-12, SaCV-7, DflaCV-1, DflaCV-2, LDMD-12, LDMD-18, RW-E and SI00898, while the rest all have a bidirectional genome organisation (Fig. 8.1). Interestingly, RW-E is from a treated sewage sample collected in Florida, USA at point of discharge into environment.

The Reps of SaGmV-1 to -11 are most closely related to those of gemycircularviruses (Fig. 8.3) with SaGmV-4 being the most divergent of these viruses. The Rep sequences of four SaGmVs (SaGmV-1a, SaGmV-7a, SaGmV-7b, SaGmV-8, and SaGmV-9) form a well-supported clade together with BasCV-3, with sequences in the clade sharing >76.45% Rep aa identity. Interestingly, the Rep of SaGmV-6 shares 70.19% pairwise identity with SsHADV-1 Rep sequences recovered in China and New Zealand (Kraberger *et al.*, 2013; Yu *et al.*, 2010). Given that SsHADV-1 infects the fungus *Sclerotinia sclerotiorum* it is possible that other members of this clade also infect fungi. Hypericum japonicum-associated circular DNA virus (HJasCV) (Du *et al.*, 2014) and Cassava-associated circular DNA virus (CasCV) (Dayaram *et al.*, 2012) were isolated from plant samples and therefore may infect epiphytic/endophytic fungi living on or within plants. Since several viruses in this group were isolated from faecal samples it is possible these originated from fungal material or spores consumed by the

various animals and shed in their faecal material or fungal contamination post-shedding (Ng *et al.*, 2014; Sikorski *et al.*, 2013b). Those gemycircularviruses associated with dragonflies (Rosario *et al.*, 2012a) and mosquitos (Ng *et al.*, 2011) may be from fungal material indirectly consumed by the dragonfly or adhered to the outside of the insect. Recently the discovery of gemycircularviruses in bovine serum from healthy cattle as well as from brain tissue and serum of a patient with multiple sclerosis perhaps adds a new dimension on the possible host range of these viruses (Lamberto *et al.*, 2014). SaGmV-10a and 10b are closely related, sharing 81.1% and 87.5% genome-wide pairwise identity respectively to the gemycircularvirus isolate recovered from serum and brain tissue from a patient with multiple sclerosis (isolates MSSI2.225 and MSBI3.224, which are said to be identical).

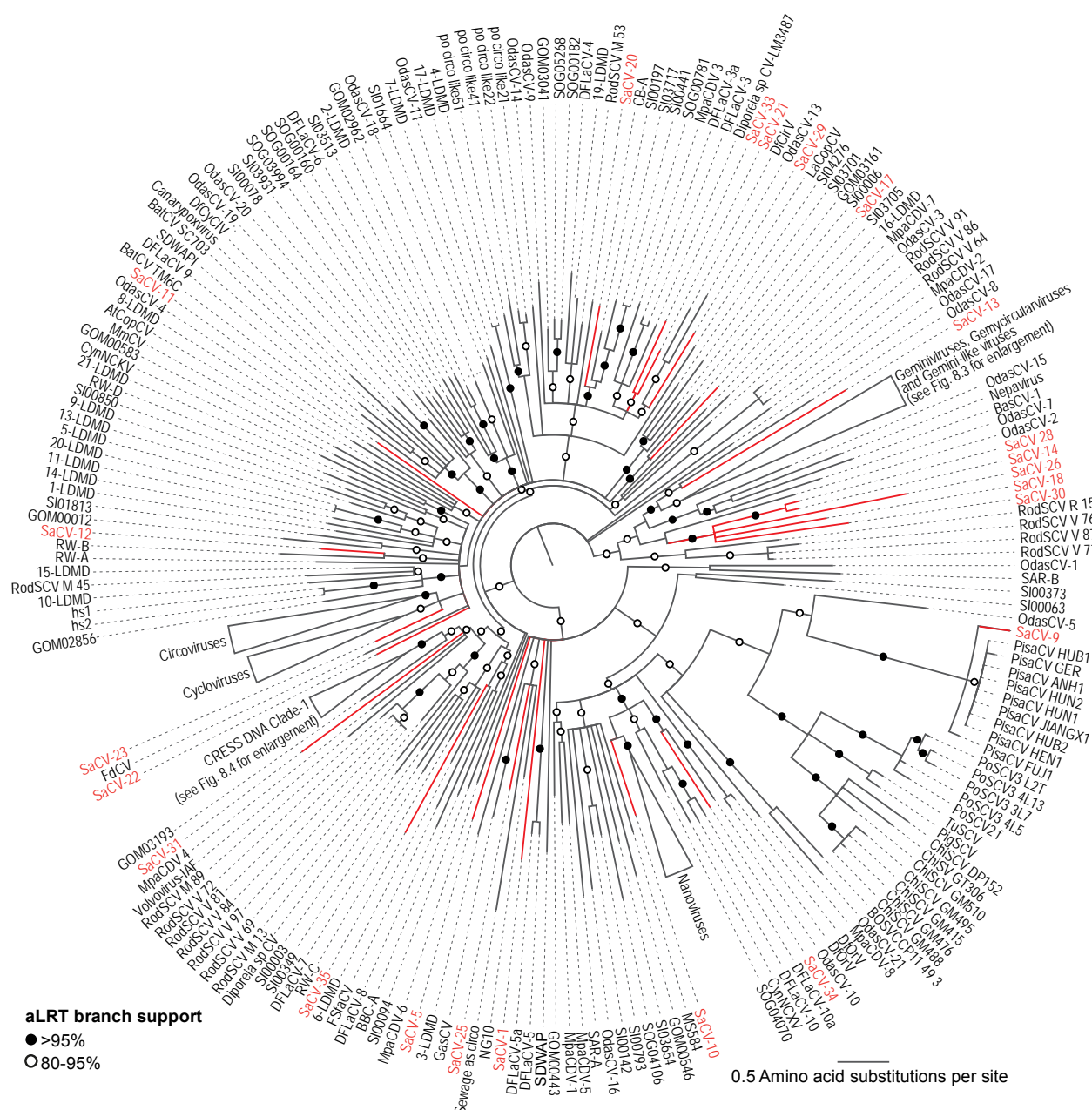


Figure 8.2: Mid-point rooted maximum likelihood phylogenetic tree (constructed with the nucleotide substitution model WAG+G) of Rep amino acid sequences encoded by known and putative CRESS DNA viral genomes recovered in this study. Phylogenetic tree was midpoint rooted. Branches with <80% aLRT support have been collapsed. Accession numbers for the CRESS DNA viruses are provided in Table 4. Branches and names in red indicate Reps encoded by viral genomes recovered in this study. The clade for the Nanoviruses, cycloviruses, circoviruses, CRESS DNA clade-1 and major grouping incorporating the geminiviruses, gemycircularviruses, and gemini-like viruses have been collapsed.

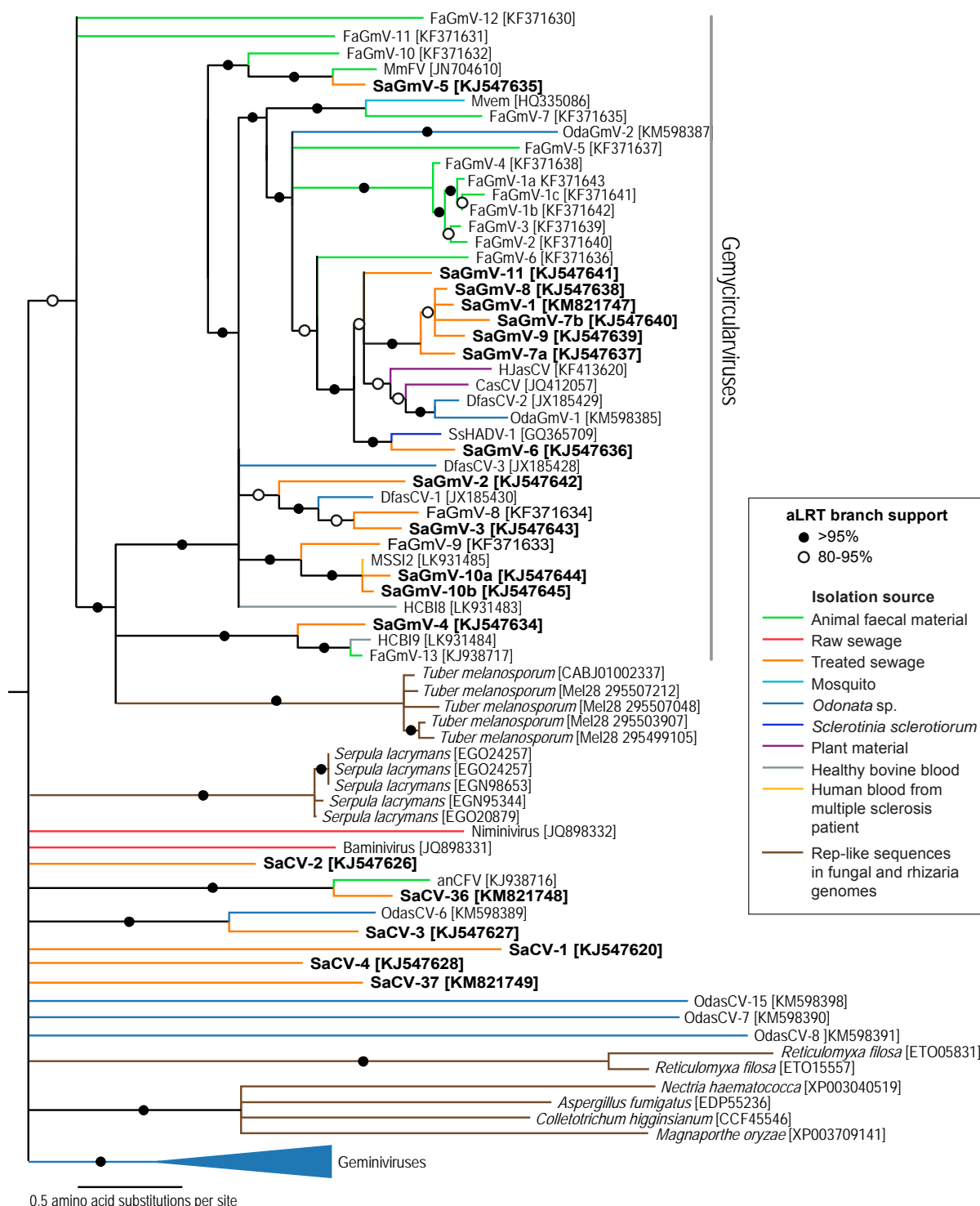


Figure 8.3: Maximum likelihood Rep amino acid sequence phylogenetic trees (constructed with the nucleotide substitution model LG+I+G+F). Sources of viral isolates are indicated by colours shown in the key. Branches with aLRT <80% support have been collapsed. Reps encoded by CRESS DNA viruses recovered in this study are indicated in bold. Mid-point rooted maximum likelihood phylogenetic tree of all known gemycircularvirus Rep sequences including those from this study.

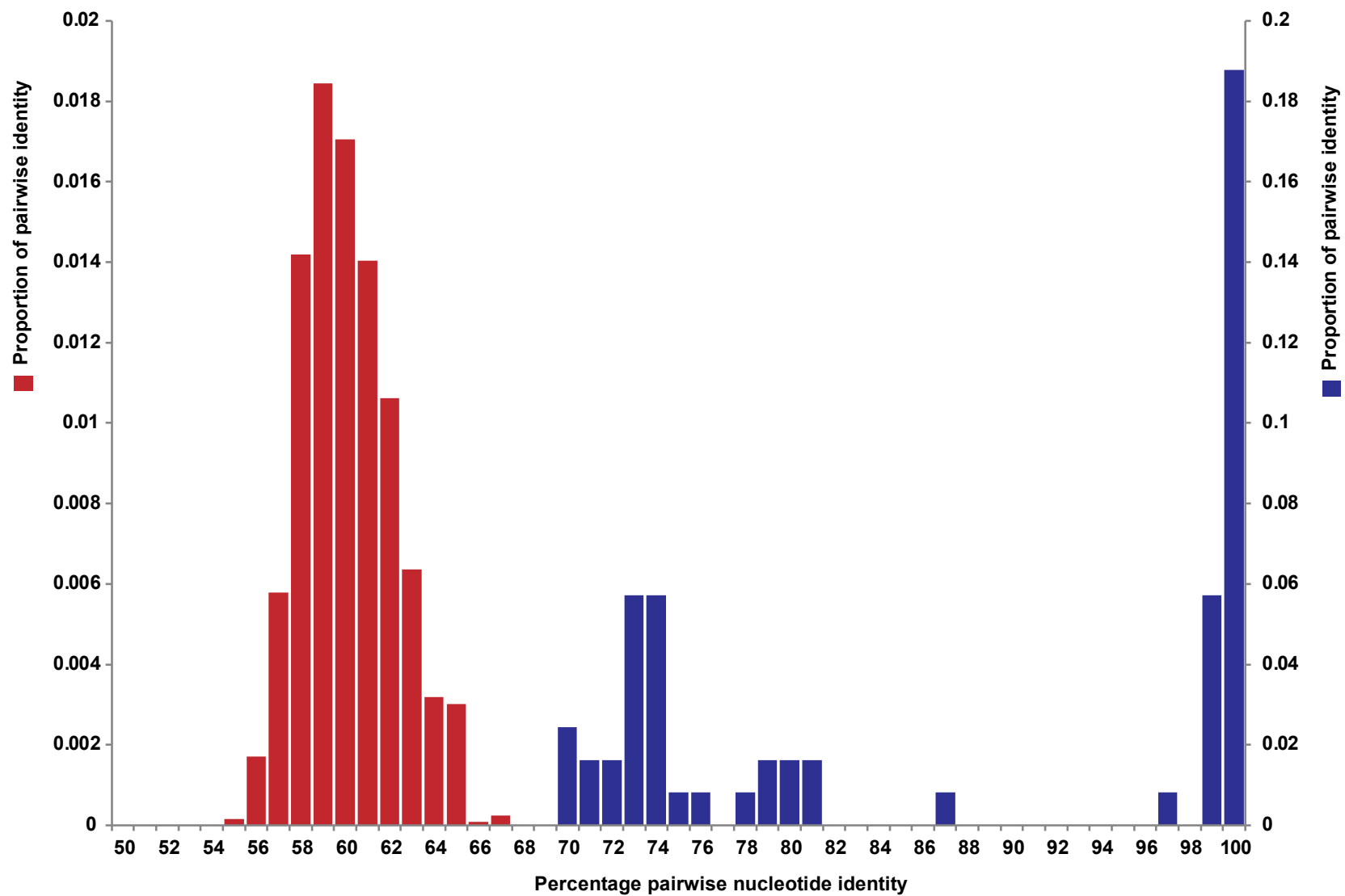


Figure 8.4: Distribution of genome-wide pairwise nucleotide identities of gemycircularviruses.

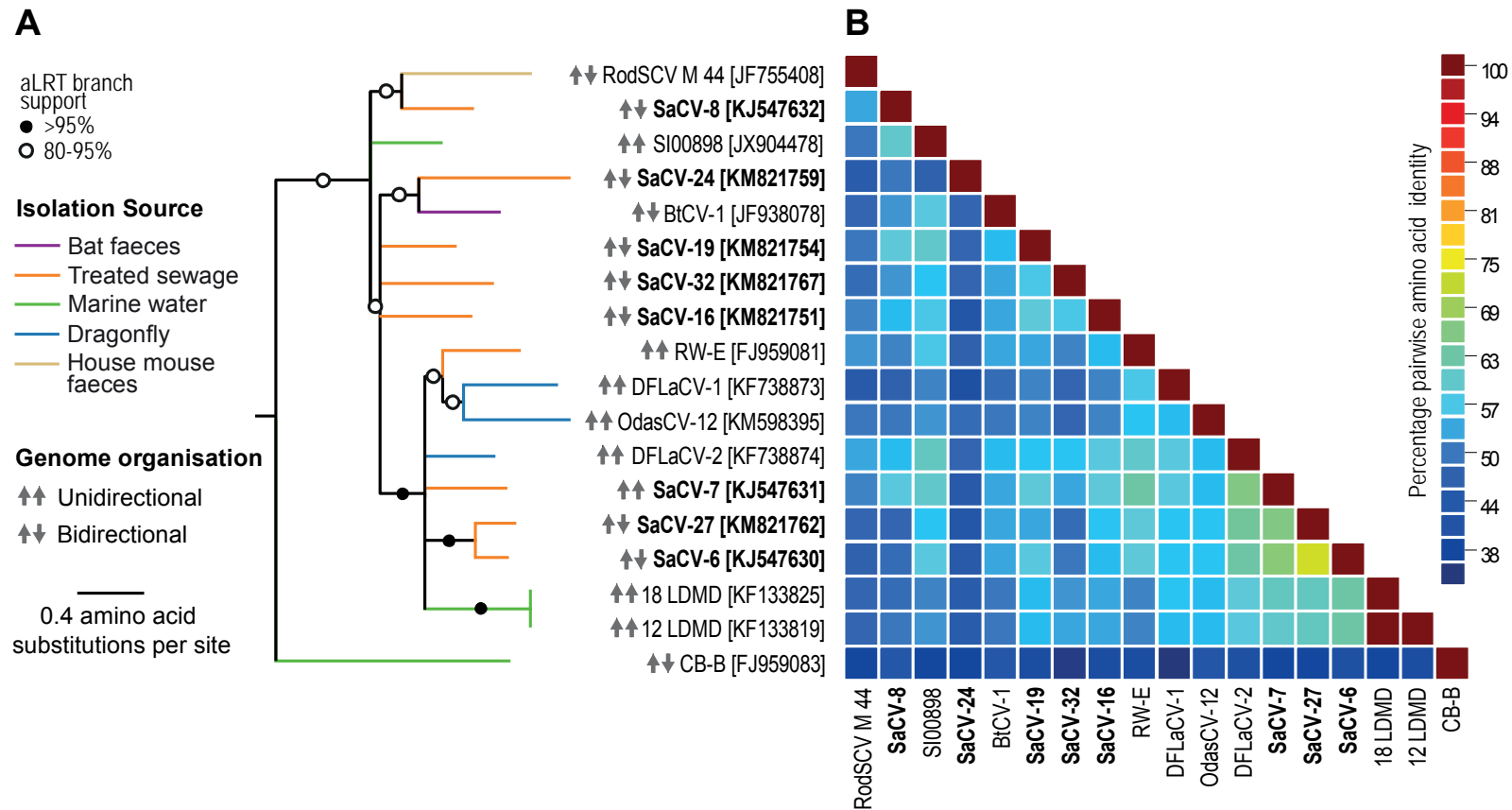


Figure 8.5. Maximum likelihood Rep amino acid sequence phylogenetic trees (constructed with the nucleotide substitution model LG+I+G+F). Maximum likelihood phylogenetic tree (A) and pairwise amino acid comparison matrix (B) of Rep sequences from CRESS DNA virus Clade 1. CRESS DNA viruses recovered in this study are indicated in bold.

8.4.3 Conserved motifs within replication-associated proteins

All well studied circular ssDNA viruses that express a Rep protein replicate by a rolling-circle mechanism (Jeske *et al.*, 2001). There are a number of well conserved motifs present in the Reps of these viruses (Koonin & Ilyina, 1992; Londoño *et al.*, 2010; Rosario *et al.*, 2012b) that are characteristic of rolling-circle replication (RCR) initiator proteins. Besides facilitating the definitive identification of Rep-encoding ORFs, the extreme conservation of these motifs is the primary reason that Reps expressed by diverse rolling-circle replicons have identifiable homology. The most conserved of these motifs, known as RCR motifs I, II and III have been reviewed in Rosario *et al.* (2012b). A fourth large conserved motif known simply as the geminivirus Rep sequence (GRS) domain has been identified in geminiviruses (Nash *et al.*, 2011) and gemycircularviruses (Dayaram *et al.*, 2012; Sikorski *et al.*, 2013b). Three putative conserved helicase domains known as Walker A, B and C motifs sit downstream of motif III, these have been identified in all the major families of eukaryote-infecting circular ssDNA viruses (Rosario *et al.*, 2012b).

We attempted to identify these various motifs within all the probable Rep sequences encoded by the genomes recovered in this study (Table 3 and 4). The Reps of gemycircularviruses and those closely related have motifs (including the GRS motif) which are most similar to those found in geminivirus Reps (Table 3). Interestingly, Niminivirus and SaCV-4 both have the same motif I sequence (FLTYPQ) as that found in most geminivirus Reps (Table 3). Motif I has been shown in geminiviruses to be required for DNA binding and cleavage prior to RCR (Orozco & Hanley-Bowdoin, 1998) and it likely has a similar function in these other viruses. All the SaCVs and the SaGmVs have a conserved motif III sequence (YxxK), with the exception of SaCV-10, SaCV-34 and SaCV-35 (Table 3 and 4), that is similar to that seen in all the well characterised eukaryotic ssDNA viral families. There is a high degree of conservation among the members of CRESS DNA Clade 1 viruses, especially in motif I [(M/L)LT(A/I)P], motif III [YV(W/G)K] and the Walker C motif [(F/I)TSN] (Table 2).

Table 8.3: Putative conserved motifs identified in the Reps of gemycircularviruses and those that are most closely related.

Viral isolate	Genbank accession #	RCR motif I	RCR motif II	GRS motif	RCR motif III	Walker-A	Walker-B	Motif C
SaGmV-1	KM821747	LITYSQ	VHLHC	RCFDIEGRHPNIEPSR	YACK	GESRTGKT	AIFDDW	VISN
SaGmV-2	KJ547642	IVTYAQ	IHLHC	DVFDVEGRHPNVQHV	YCIK	GPTRVGKT	VFDDM	YCHN
SaGmV-3	KJ547643	LLTYAQ	LHIHA	RVFDMGCHPNIVRGY	YAIK	GPTKLGKT	VFDDM	YLYN
SaGmV-4	KJ547634	IVTFPQ	VHYHM	NLFDYFGAHGNISVR	YCGK	GPSRTGKT	VFDDI	MCLN
SaGmV-5	KJ547635	LLTYSQ	IHLHA	RRFDVEGYHPNIQPCG	YAIK	GETRLGKT	VLDDI	WLMN
SaGmV-6	KJ547636	LLTYAQ	IHLHV	DTFDVGGFHPNISQSY	YACK	GPSQTGKT	VFDDI	WCSN
SaGmV-7a	KJ547637	LLTYAQ	VHLHV	NIFDVGCHPNISPS	YAIK	GESRTGKT	VFDDI	WLSN
SaGmV-7b	KJ547640	LLTYSQ	VHLHC	RCFDIEGRHPNVEPSH	YVIK	GESRTGKT	VFDDV	WLSN
SaGmV-8	KJ547638	LITYAQ	VHLHC	RVFDIEGRHPNIEPSR	YAVK	GESRTGKT	VFDDI	WLSN
SaGmV-9	KJ547639	LLTYAQ	IHLHC	RVFDIENRHPNIEPSR	YAVK	GESRTGKT	VFDDI	WLSN
SaGmV-10a	KJ547644	LLTYAQ	IHLHA	RAFDVEGQHPNVSPSR	YAIK	GPSRMGKT	IFDDF	WLSN
SaGmV-10b	KJ547645	LLTYPQ	LHLHA	RAFDVEGCHPNVSPSR	YAIK	GPSRMGKT	IFDDF	WLSN
SaGmV-11	KJ547641	LLTYAQ	LHLHV	DLFDVDGHHHPNVTPSR	YAIK	GRSRTGKP	VFDDI	WLAN
FaGmV-1a	KF371643	LLTYAQ	THLHA	DVFDVGGRRHPNLVPSY	YAIK	GDTRLGKT	VFDDM	WLAN
FaGmV-1b	KF371642	LLTYAQ	THLHA	DVFDVGGRRHPNLVPSY	YAIK	GDTRLGKT	VFDDM	WLAN
FaGmV-1c	KF371641	LLTYAQ	THLHA	DVFDVGGRRHPNLVPSY	YAIK	FPAWLDVV	AIDDM	WLAN
FaGmV-2	KF371640	LLTYAQ	THLHA	DVFDVGGRRHPNVMPF	YATK	GDTRLGKT	VFDDM	WLSN
FaGmV-3	KF371639	LLTYAQ	THLHA	DVFDVGGRRHPNLVPSY	YAIK	GDTRLGKT	VFDDM	WLSN
FaGmV-4	KF371638	LLTYAQ	THLHA	DVFDVGGFHPNIEASR	YAIK	GDTRLGKT	VFDDM	WLAN
FaGmV-5	KF371637	LVTYPQ	THLHV	DIFDVGGFHPNIESK	YACK	GDALTGKT	VIDDI	WIAN
FaGmV-6	KF371636	LLTYAQ	IHLHC	RIFDVGRRHPNVVPSR	YAIK	GPSLTGKT	VLDDI	WCAN
FaGmV-7	KF371635	LLTYPQ	YTSHC	RIFDIQGHHPNIEVRG	YTIK	GETRLGKT	IFDDL	WCSN
FaGmV-8	KF371634	LLTFPQ	LHLHA	RVFDVGGRRHPNVVRGY	YAIK	GPTRLGKT	IFDDM	YICN
FaGmV-9	KF371633	LLTYAQ	IHLHA	RVFDVQGHHPNVEPSR	YAIK	GPTRTGKT	VFDDF	WINN
FaGmV-10	KF371632	CPHYLP	THLHA	RRFDVDGYHPNVQPF	YAIK	GESRLGKT	VFDDM	WLCN
FaGmV-11	KF371631	FLTYSQ	HHYHV	RIFDVGGCHPNFKSVR	YCLK	GRSRLGKT	VMDDI	WCTN
FaGmV-12	KF371630	FLTYSQ	FHFHA	RIFDFDLGHPNIESVR	YTKK	GPHRRRT	VFDDI	WVCN
FaGmV-13	KJ938717	IITFPQ	IHYHV	DSFDVLGHHHPNWTPIR	LEHG	GPTRTGKT	VFDDI	MCMN
HCB19	LK931484	ITFPQV	IHYHI	TAFDYFGAHGNISIR	YVGK	GPTRTGKT	VFDDI	MCMN
HCB18	LK931483	LTYAQC	THLHA	AVFDVGGFHPNISITK	YAIK	GPSRTGKT	VFDDI	WISN
MSSI2	LK931485	LTYPQC	LHLHA	RAFDVEGCHPNVSPSR	YAIK	GPSRMGKT	IFDDF	WLSN
MmFV	JN704610	LLTYAQ	IHLHA	RRFDVEGFFHPNIAPCG	YAIK	GETRLGKT	VLDDM	WLMN
MVemV	HQ335086	LLTYAQ	IHFHA	RFWDIAGRHPNIARVG	YAIK	GPSRTGKT	VFDDI	WVSN
SsHADV-1	GQ365709	LLTYAQ	IHLHC	DVFDVDGHHHPNITKSR	YAIK	GPSQTGKT	VFDDI	WCSN
CasCV	JQ412057	LITYAQ	VHLHC	DIFDVGRRHPNIEPSW	YAIK	GDSRSRGT	IFDDI	WISN
HjasCV	KF413620	LVTYAQ	LHLHV	DILDVDGRRHPNLAPIK	YAIK	GGTRTGKT	VFDDI	WICN
DfasCV-1	JX185430	LLTYPQ	VHLHA	RVFDVDGHHHPNIVRGY	YATK	GDTRLGKT	VFDDM	YISN
DfasCV-2	JX185429	LVTYPQ	LHLHC	DIFDVGCHPNIQPST	YAIK	GESRTGKT	IFDDI	WISN
DfasCV-3	JX185428	LLTYAQ	THYHA	RIFDIDGYHPNLSGR	YATK	GPSRTGKT	VFDDI	WCNN
DfasCV-4	KM598385	LITYAQ	LHLHV	DIFDVGRRHPNKRWS	YAIK	GGNGSGQT	IFDDI	WLCN
DfasCV-5	KM598387	LLTYSQ	THFHV	NVFDVGGHHHPNLPVW	YAAK	SSLAFRKP	VFDDW	WLCN
SaCV-1	KJ547620	IITYPQ	LHRHA	FFDHLTRHPNICKVGK	YVKK	GASRLGKT	VFDDI	TSTN
SaCV-2	KJ547626	FLTYPR	LHVHA	RFFDVAGFHPNIQTVR	YLDK	GPSRTGKT	IFDDV	VDN
SaCV-3	KJ547627	LLSEQN	LHLHA	DAFDVDGFFHPNIQKPR	YCSK	GKSRWGKT	IFDDI	WLCN
SaCV-4	KJ547628	FLTYPQ	PHLHA	TFFNYENYHPNIQSAR	YTKK	GPSKLK	VFDDF	YCAN
NimiV	JQ898332	FLTYPQ	PHFHA	RHFDISGYHPNIQVCR	YVTK	GPSKTGKS	VLDDI	IVCS
BamiV	JQ898331	LLTYPQ	PHLHI	RFFDVTDFHPNVVVVR	YIAK	GASRIGKT	VFDDI	FLVN
SaCV-36	KM821748	FLTYSQ	IHYHV	DVFDLDNHHHPNIAIIK	YIRK	GPTRLGKS	ILDDF	WICQ
SaCV-37	KM821749	IITYPR	PHIHV	RYFDIGDHHHPNVQSTR	YVAK	GELLYIVT	VFDDI	WLSN

Table 8.4: Putative conserved motifs identified in the Repts encoded by CRESS DNA viruses identified in this study and of those of CRESS DNA virus Clade 1.

Viral isolate	Genbank accession #	RCR motif					
		I	II	III	Walker-A	Walker-B	Motif C
SaCV-5	KJ547629	FYTYNN	PHLQG	YCTK	GPGGVGKD	ILDEF	ILTN
SaCV-9	KJ547633	CEGDNA	RHFQF	YVYK	ERGNSGKT	IIDTP	VLCN
SaCV-10	KJ547621	CFTHNN	DHIQG	YCME	GNNGKTYF	IFDYV	VLAN
SaCV-11	KJ547622	VFTLNN	PHLQG	YCTK	GPTGAGKT	ILDDI	ITSN
SaCV-12	KJ547623	CFTLNN	PHLQG	YCHK	GVTGTGKT	VIDDF	ITSN
SaCV-13	KJ547624	LCTYAK	KHMHV	YL GK	GNSWKS YT	FVDLA	VLAN
SaCV-14	KJ547625	FLTWPQ	KHVHA	YVCK	GLPGVGKT	VLDEY	IVSN
SaCV-15	KM821750	SITINN	VHYQG	YVHK	GRWVSTEF	YVDSL	ILRH
SaCV-17	KM821752	CFTLNN	RHLQG	YCSK	GLPGVGKS	IIDDF	VTSN
SaCV-18	KM821753	TITFPQ	PHLHV	YCTK	GPKNLGKT	VFDEF	ILSN
SaCV-20	KM821755	VFTVNN	PHLQG	YCKK	GKAGCGKS	IIEDF	ITSN
SaCV-21	KM821756	CFTLNN	RHLQG	YCTK	GEPGVGKS	ILDDF	VTSN
SaCV-22	KM821757	CFTQNN	PHYQG	YCTK	GPPGTGKS	VIDEF	ITSN
SaCV-23	KM821758	VFTVNN	PHIQG	YCRK	GPPGSGKS	IIDDF	ITSN
SaCV-25	KM821760	VFTVNN	PHIQG	YCTK	GPTGTGKT	VLEEF	ITSN
SaCV-26	KM821761	SLTYPQ	VHRHV	YCMK	EAPNLGKT	LLDEY	ITSN
SaCV-28	KM821763	FLTWPQ	PHVHA	YVCK	GLPGVGKT	VLDEF	IVSN
SaCV-29	KM821764	CYTLNN	PHHQG	YCKK	GASGAGKT	IIDDV	ITSQ
SaCV-30	KM821765	FLTFPQ	KHLHA	YVMK	GPPGIGKT	LLDEF	IMSN
SaCV-31	KM821766	CVTWNN	PHFQM	YCTK	GPSAVGKT	IVDEW	FTSN
SaCV-33	KM821768	LLTIND	EHWQL	YVFK	GPPGVGKS	ILDDF	ITSN
SaCV-34	KM821769	CFTSFN	KHIQG	YCKQ	MGGNTGKS	IYDLA	VMAN
SaCV-35	KM821770	IGTIYL	HHIQI	YCTE	GGAGVGKS	LFDDF	FTAD
SaCV-36	KM821748	FLTYSQ	IHYHV	YIRK	GPTRLGKS	ILDDF	WICQ
CRESS DNA virus clade-1							
SaCV-6	KJ547630	LLTIPH	KHWQI	YVWK	GPTGTGKS	VIDEF	ITSN
SaCV-7	KJ547631	LLTIPH	LHWQA	YVWK	GRTGTGKS	VIDEF	ITSN
SaCV-8	KJ547632	MLTAPA	SHWQL	YVWK	GRTGTGKS	VIDEF	ITSN
SaCV-16	KM821751	ILTIPH	LHWQL	YVWK	GPTGVGKS	VIDEF	ITSN
SaCV-19	KM821754	ILTIPH	LHWQI	YVHK	GRTGTGKS	VIDEF	ITSN
SaCV-24	KM821759	IGTISI	QHWQV	YVWK	GLTGTGKS	VIDEF	ITSN
SaCV-27	KM821762	LLTIPF	LHWQI	YVWK	GPTATGKS	VLDEY	ITTN
SaCV-32	KM821767	LLTIPH	LHWQL	YVWK	GDAGTGKS	VIDEF	ITSN
DFLaCV-1	KF738873	CFTVNN	SHWQI	YVWK	GDTRLGKT	IVDEF	FTSN
DFLaCV-2	KF738874	VFTLNN	LHWQV	YVWK	GESRTGKT	VIDEF	ITSN
RWE	FJ959081	LLTIPE	LHWQV	YVWK	GPTGTGKS	VVDEF	ITSN
CB-B	FJ959083	ILTIPE	RHWQV	YVGK	GPTGTGKT	IIDEF	ITSN
RodSCV-M44	JF755408	MLTIPY	PHWQL	YVWK	GSTGMGKS	VIDEF	ITSN
BiCV-1	JF938078	LLTIPY	LHWQL	YVWK	GRTGAGKS	VIDEF	ITSN
SI00898	JX904478	LLTIPH	LHWQI	YVWK	GRTETGKS	VMDEF	ITSN
12-LDMD	KF133819	MLTIPH	IHWQL	YVWK	GVSGSGKS	VLDEF	ITSN
18-LDMD	KF133825	MLTIPH	IHWQL	YVWK	GVSGSGKS	VLDEF	ITSN
OdasCV-12	KM598395	LGTIPE	EHWQV	YVWK	GPTGTGKS	VIDEF	ITSN

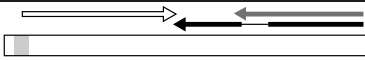
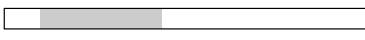
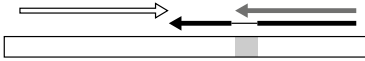


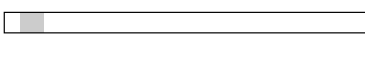
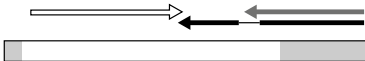
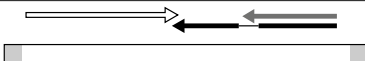
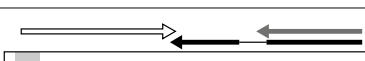
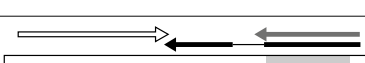
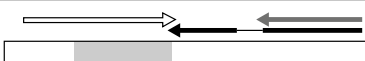
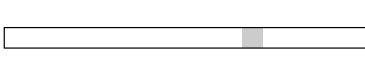
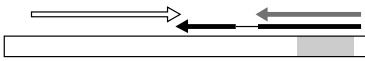
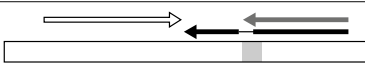
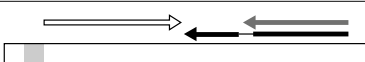
8.4.4 Nonanucleotide sequence analysis

In order to initiate RCR the Rep of ssDNA viruses in the family's *Circoviridae*, *Nanoviridae* and *Geminiviridae* cleaves the parental virion-sense DNA strand within a conserved nonanucleotide motif in the loop domain of a stem-loop structure that is usually located in an intergenic region. Probable virion-strand origin of replication nonanucleotide sequences were identified in all 50 of the complete genomes recovered in this study (Fig. 8.1). All gemycircularviruses, with the exception of three genomes (MmFV, FaGmV-12, -13 and SaGmV-5) have a TAATRTTAD nonanucleotide sequence, whereas the complete genomes with Rep sequences that cluster close to the gemycircularviruses (Baminivirus, Nimivirus, SaCV-2, -3, -4 and SaCV-37) have a TAATATTAC nonanucleotide sequence, which is highly conserved amongst the geminiviruses (the primary exceptions being *Eragrostis curvula streak virus* [FJ665631] (Varsani *et al.*, 2009b) and *Beet curly top Iran virus* [EU273816] (Yazdi *et al.*, 2008)). The nonanucleotide motifs of SaCV-5 through -14 are variable (Table 2) and to some extent this is not surprising given how divergent these are from the other viral genomes recovered in this study (Fig. 8.2; Additional Table 8.1).

8.4.5 Recombination patterns among Gemycircularviruses

Of all the complete genomes recovered in this study, only those that were closely related to the gemycircularvirus group could be aligned with sufficient accuracy to enable reliable recombination analysis. Previously Sikorski *et al.* (2013b) isolated 12 new gemycircularviruses and analysed these along with all other previously identified gemycircularviruses for evidence of recombination (Fig. 8.6). Their results indicated seven strongly supported recombination events were detected. Given that we have significantly increased the number of known viral genomes within this group we were able to identify a total of 15 recombination events. Although the addition of several new sequences to this group has increased the alignment credibility on a full genome scale, there is still a relatively high degree of sequence diversity and therefore it is important to keep in mind that the position of individual breakpoints may vary as sequences that are more closely related to the parental viruses of recombinants are recovered. The fragments of sequence that were apparently transferred during recombination vary in size from 90nt (event 1 in Fig. 8.6) to 786nt (event 5 in Fig. 8.6). Breakpoints were found spanning coding and non-coding regions with the exception of event 8 which was exclusively limited to non-coding regions (Fig. 8.6).

There are currently no obvious recombination breakpoint hot- or cold-spots within the gemycircularvirus genomes. This is in contrast to observations made in very well sampled ssDNA virus families such as the circoviruses, nanoviruses and geminiviruses. In these other virus groups whereas recombination breakpoint hotspots tend to occur in intergenic regions and are particularly common at the virion-strand origin of replication (Kraberger *et al.*, 2012; Lefeuvre *et al.*, 2007; Martin *et al.*, 2011b; Varsani *et al.*, 2009a), recombination breakpoint cold-spots tend to occur within the internal regions of coat protein genes. Recombination hotspots are thought to occur in the intergenic regions because this is where replication and transcription factors may clash during replication (Martin *et al.*, 2011a).

Event	Recombinant(s)		Recombinant region	Potential minor parent	Potential major parent	Detection method	p-value
1	FaGmV-1b (KF371642) FaGmV-1c (KF371641)		62–151	Ancestral FaGmV-5-like (KF371637)	FaGmV-1a (KF371643)	RGBMCST	4.14 x10 ⁻¹⁷
2	FaGmV-1b (KF371642) FaGmV-1c (KF371641)		218–990	Ancestral FaGmV-5-like (KF371637)	FaGmV-1a (KF371643)	RBMC	2.07 x10 ⁻¹⁶
3	SaGmV-4 (KF371638)		1464–1605	Ancestral FaGmV-5-like (KF371637)	FaGmV-3 (KF371639)	RMCST	4.83 x10 ⁻¹³
4	DfaCV-2 (JX185429)		1611–2133	HjasCV (KF413620)	FaGmV-5 (KF371637)	RBMCST	1.07 x10 ⁻¹⁹
5	ssHADV-1 (GQ365709, KF268025, KF268026, KF268027, KF268028, KM598382, KM598383, KM598384)		245–1030	FaGmV-1a (KF371643)	Ancestral FaGmV-1c-like (KF371641)	RBMCT	3.31 x10 ⁻²⁹
6	ssHADV-1 GQ365709, KF268025, KF268026, KF268027, KF268028, KM598382, KM598383, KM598384)		100–244	FaGmV-5 (KF371637)	SaGmV-6 (KJ547636)	RMS	6.69 x10 ⁻⁰⁷
7	SaGmV-7b (KJ547640)		1698–109	SaGmV-9 (KJ547639)	SaGmV-7a (KJ547637)	RGS	8.05 x10 ⁻⁰⁷
8	FaGmV-6 (KF371636)		2188–104	FaGmV-1a (KF371643)	Ancestral FaGmV-5-like (KF371637)	MCS	8.74 x10 ⁻¹⁴
9	FaGmV-5 (KF371637)		70–212*	FaGmV-1a (KF371643)	Ancestral FaGmV-2-like (KF371640)	RGBMC	1.36 x10 ⁻⁰⁶
10	HjasCV (KF413620)		1587–2096	SaGmV-11 (KJ547641)	FaGmV-1a (KF371643) FaGmV-1c (KF371641) FaGmV-3 (KF371639)	RBMCS	3.25 x10 ⁻¹⁰
11	FaGmV-3 (KF371639)		422–1018	DfaCV-2 (JX185429)	FaGmV-1a (KF371643)	RMC	6.08 x10 ⁻⁰⁸
12	FaGmV-3 (KF371639)		1442–1566	FaGmV-5 (KF371637)	FaGmV-1b (KF371642) FaGmV-1c (KF371641) FaGmV-2 (KF371640)	RGBMC	5.64 x10 ⁻¹²
13	SaGmV-7a (KJ547637)		1852–2205	DfaCV-2 (JX185429) HjasCV (KF413620)	SaGmV-7b (KJ547640) SaGmV-8 (KJ547638)	RGBMCST	8.36 x10 ⁻¹⁷
14	SaGmV-10b (KJ547645)		1506–1628	SaGmV-10a (KJ547644)	MSSI2 (LK931485)	RGBMCT	2.90 x10 ⁻¹⁰
15	MSSI2 (LK931485)		124–299	SaGmV-10a (KJ547644)	MSSI2 (LK931485)	RBMC	6.75 x10 ⁻⁰⁷

RDP (R) GENCONV (G), BOOTSCAN (B), MAXCHI (M), CHIMERA (C), SISCAN (S) and 3SEQ (T)

cp↷*repA*➡*rep*➡■ recombinant region

* = The actual breakpoint position is undetermined.

Figure 8.6: Illustration and details of recombination events detected in all gemycircularvirus genomes. Arrows above genome maps indicate the positions of large ORFs. Grey boxes represent recombination event. Major and minor parents indicate the most likely identities of parental sequences that respectively donated the larger and smaller regions of the recombinants genome. Methods used to detect recombination are as follows RDP (R), GENCONV (G), BOOTSCAN (B), MAXCHI (M), CHIMERA (C), SISCAN (S) and 3SEQ (T). For each event the method with the most significant associated p-value is indicated in bold.

8.5 Concluding remarks

Prior to the advent of metagenomic sequencing of CRESS DNA viruses the diversity of this group was largely underestimated. It now seems apparent that such viruses are likely ubiquitous in the biosphere and may collectively represent a more common biological entity on Earth than previously thought (Londoño *et al.*, 2010; López-Bueno *et al.*, 2009; McDaniel *et al.*, 2013; Rosario *et al.*, 2009b; Roux *et al.*, 2012; Zawar-Reza *et al.*, 2014).

This study investigates CRESS DNA viral diversity of circular genomes recovered from an oxidation pond using a high-throughput sequencing-informed approach. Using back-to-back primers based on Illumina derived *de novo* contig assemblies, the recovery and Sanger sequencing of full genomes both enables verification of Illumina-derived sequencing contigs present in the sample, and allows the recovery of full genomes for future biological characterisation. We postulate that while many of the 50 putatively complete genomes recovered here likely represent viruses that infect a variety of microorganisms within oxidation ponds, some of these genomes may also have been derived from human excrement and might therefore represent viruses infecting humans, their gut-associated microflora or their food-sources, such as plants. In order to identify the uni- or multi-cellular species within the environment in which these viruses either replicate or are transported, the sequences presented could be used to produce specific probes to detect closely related homologues within a variety of organisms. For example, through environmental sampling of faecal matter and insects a novel viral group named Cyclovirus (Dayaram *et al.*, 2013; Ge *et al.*, 2011; Li *et al.*, 2010a; Rosario *et al.*, 2012a; Rosario *et al.*, 2011) was discovered: This group is closely related to circoviruses and possibly represents a novel genus within the family *Circoviridae*. Over the last year cycloviruses have been found in human cerebrospinal fluid and nasopharyngeal aspirates (de Jong *et al.*, 2014; Phan *et al.*, 2014; Smits *et al.*, 2012; van Doorn *et al.*, 2013). Therefore, baseline data of diverse CRESS DNA viruses collected using metagenomic approaches from ecosystems can prove very useful in the identification of related viruses in various organisms.

Gemycircularviruses have been recovered from a wide range of different organisms and environmental sources (Dayaram *et al.*, 2012; Du *et al.*, 2014; Kraberger *et al.*, 2013; Ng *et*

al., 2014; Rosario *et al.*, 2012a; Sikorski *et al.*, 2013b; Yu *et al.*, 2010), suggesting that these viruses have a very broad geographical distribution and are present in a variety of different ecosystems. SsHADV-1 is the only gemycircularvirus for which a host species has been definitively identified. This virus is known to infect the fungus, *Sclerotinia sclerotiorum* and it has therefore been proposed that other viruses within this group may also infect fungi (Dayaram *et al.*, 2012; Sikorski *et al.*, 2013b). Further evidence that these viruses may infect fungi is the discovery of gemycircularvirus-like Rep sequences integrated within the genomes of various fungal species. Further, a large number of fungal species have been identified in sewage (Becker & Shaw, 1955; Dorcas *et al.*, 2013; Ismail & Abdel-Sater, 1994; Kacprzak *et al.*, 2005; Ulfig *et al.*, 1996) and it is therefore entirely plausible that the SaGmVs identified here may infect either fungi growing in/around the oxidation ponds, or fungi associated with excrement itself. There is also plentiful algal growth in oxidation ponds (Abdel-Raouf *et al.*, 2012; Oswald *et al.*, 1953) and the possibility remains that algae may also be the hosts of some of the identified viruses. We do however reiterate the earlier comment regarding the discovery of gemycircularviruses associated with a multitude of sources including serum and brain tissue leaving the question wide open as to the host range of these viruses. The identification of recombination events among the gemycircularviruses highlights that at least some of these viruses must have an overlapping host range in order for recombination to occur.

Although there is generally a reduction in the titres of human pathogenic viruses following sewage treatment many, such as rotaviruses and adenoviruses persist and remain detectable both in treated sewage (Hewitt *et al.*, 2011; Tonani *et al.*, 2013), and in the waterways and coastal waters into which treated sewage is pumped (Ming *et al.*, 2014; Schlindwein *et al.*, 2010; Sdiri-Loulizi *et al.*, 2010; Van Heerden *et al.*, 2003). Many sewage-associated viruses are in fact so persistent that they can often be found within filter feeders living on coastlines near sewage outflows (Kittigul *et al.*, 2014; Seo *et al.*, 2014). Such studies have shown that many viruses are able to withstand the sewage treatment process highlighting the importance of identifying the overall viral diversity present in treated sewage in order to have an awareness of what viral populations are discharged into the ocean or waterways. Although the viruses identified in this study most likely pose no risk to public health, they are nevertheless important in extending our current knowledge both on the biodiversity at the interface between terrestrial and aquatic viromes, and on ssDNA virus diversity in the general

environment. While our study provides baseline data on ssDNA viral diversity in treated sewage, future metagenomics studies investigating viral populations in raw sewage to those in treated sewage would give us a better indication as to whether these viruses originate in human excrement or are introduced during the open-air stage of sewage treatment. A significant portion of the viruses isolated from treated sewage share similarity to geminiviruses, which is the only clue we have to their possible hosts. Further research is needed to identify if in fact these gemini-like viruses do infect plants or other organisms.

GenBank accession numbers:

SaCV-1–37: KJ547620 – KJ547625, KM821748 – KM821770

SaGmV-1–11: KJ547634 – KJ547643, KM821747

SaCM-1–11: KJ547618 – KJ547619, KM877826 – KM877833

Additional Table 8.1: List of acronyms and corresponding accession numbers for complete CRESS DNA virus genomes available in public databases and used in this study to infer ML phylogenetic trees.

Acronym	Genbank #	Acronym	Genbank #	Acronym	Genbank #	Acronym	Genbank #
AtCopCV	JQ837277	hs2	JX559622	PoSCV 3L2T	KC5452309	SI04276	JX904605
BatCV SC703	JN857329	LaCopCV	JF912805	RodSCV M 13	JF755410	SOG00160	JX904075
BatCV TM6C	HM228875	MmCV	JQ085285	RodSCV M 44	JF755408	SOG00164	JX904076
BBC-A	FJ959086	MpaCDV-1	KJ547646	RodSCV M 45	JF755409	SOG00182	JX904077
BOSVCCP11493	JN634851	MpaCDV-2	KJ547647	RodSCV M 53	JF755415	SOG00781	JX904107
Canarypoxvirus	NP955176	MpaCDV-3	KJ547648	RodSCV M 89	JF755402	SOG03994	JX904139
CB-A	FJ959082	MpaCDV-4	KJ547649	RodSCV R 15	JF755401	SOG04070	JX904144
CB-B	FJ959083	MpaCDV-5	KJ547650	RodSCV V 64	JF755407	SOG04106	JX904147
ChiSCV DP152	GQ351272	MpaCDV-6	KJ547651	RodSCV V 69	JF755403	SOG05268	JX904185
ChiSCV GM415	GQ351277	MpaCDV-7	KJ547652	RodSCV V 72	JF755411	TuSCV	KF880727
ChiSCV GM476	GQ351274	MpaCDV-8	KJ547653	RodSCV V 76	JF755404	YNBtCV-1	JF938078
ChiSCV GM488	GQ351276	MS5845	HQ322117	RodSCV V 77	JF755405	10-LDMD	KF133817
ChiSCV GM495	GQ351273	Nepavirus	JQ898333	RodSCV V 81	JF755412	11-LDMD	KF133818
ChiSCV GM510	GQ351275	NG10	ADF80742	RodSCV V 84	JF755413	12-LDMD	KF133819
ChiSCV GT306	GQ351278	OdasCV-1	KM598393	RodSCV V 86	JF755416	13-LDMD	KF133820
CynNCKV	JX908740	OdasCV-2	KM598399	RodSCV V 87	JF755406	14-LDMD	KF133821
CynNCXV	JX908739	OdasCV-3	KM598407	RodSCV V 91	JF755417	15-LDMD	KF133822
DfCirV	JX185415	OdasCV-4	KM598408	RodSCV V 97	JF755414	16-LDMD	KF133823
DfCyCIV	JX185418	OdasCV-5	KM598410	RW-A	FJ959077	17-LDMD	KF133824
DFLaCV-1	KF738873	OdasCV-9	KM598392	RW-B	FJ959078	18-LDMD	KF133825
DFLaCV-10	KF738884	OdasCV-10	KM598412	RW-C	FJ959079	1-9LDMD	KF133826
DFLaCV-10a	KF738885	OdasCV-11	KM598394	RW-D	FJ959080	1-LDMD	KF133807
DFLaCV-2	KF738874	OdasCV-12	KM598395	RW-E	FJ959081	20-LDMD	KF133827
DFLaCV-3	KF738875	OdasCV-13	KM598396	SAR-A	FJ959084	21-LDMD	KF133828
DFLaCV-3a	KF738876	OdasCV-14	KM598397	SAR-B	FJ959085	2-LDMD	KF133808
DFLaCV-4	KF738877	OdasCV-16	KM598411	SDWAP	HQ335074	3-LDMD	KF133810
DFLaCV-5	KF738878	OdasCV-17	KM598400	SDWAPI	HQ335042	4-LDMD	KF133811
DFLaCV-5a	KF738879	OdasCV-18	KM598401	Sewage circo	ACY68125	5-LDMD	KF133812
DFLaCV-6	KF738880	OdasCV-19	KM598404	SI00003	JX904394	6-LDMD	KF133813
DFLaCV-7	KF738881	OdasCV-20	KM598406	SI00006	JX904395	7-LDMD	KF133814
DFLaCV-8	KF738882	OdasCV-21	KM598409	SI00063	JX904401	8-LDMD	KF133815
DFLaCV-9	KF738883	PigSCV	JX274036	SI00078	JX904407	9-LDMD	KF133816
DfOrV	JX185417	PisaCV ANH1	JX305997	SI00094	JX904412	Volvovirus	KC543331
DfOrV	JX185416	PisaCV FUJ1	JX305998	SI00142	JX904416		
Diporeia sp CVLM28925	KC248425	PisaCVGER2011	JQ023166	SI00197	JX904420		
Diporeia sp CV-LM3487	KC248416	PisaCV HEN1	JX305991	SI00349	JX904427		
FdCV	KC441518	PisaCV HUB1	JX305992	SI00373	JX904431		
FSfaCV	KF246569	PisaCV HUB2	JX305993	SI00441	JX904439		
GasCSV	KC172652	PisaCV HUN1	JX305995	SI00793	JX904469		
GOM00012	JX904192	PisaCV HUN2	JX305996	SI00850	JX904473		
GOM00443	JX904231	PisaCV JIANGX1	JX305994	SI00898	JX904478		
GOM00546	JX904245	pocircolike21	JF713716	SI01664	JX904518		
GOM00583	JX904250	pocircolike22	JF713717	SI01813	JX904523		
GOM02856	JX904312	pocircolike41	JF713718	SI03513	JX904541		
GOM02962	JX904333	pocircolike51	JF713719	SI03654	JX904548		
GOM03041	JX904344	PoSCV 2	KC545226	SI03701	JX904559		
GOM03161	JX904368	PoSCV 33L7	KC545227	SI03705	JX904561		
GOM03193	JX904377	PoSCV 34L13	KC545228	SI03717	JX904562		
hs1	JX559621	PoSCV 34L5	KC545229	SI03931	JX904581		

8.6 References

- Abdel-Raouf, N., Al-Homaidan, A. & Ibraheem, I. (2012). Microalgae and wastewater treatment. *Saudi Journal of Biological Sciences* **19**, 257-275.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.
- Anisimova, M. & Gascuel, O. (2006). Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Systematic Biology* **55**, 539-552.
- Becker, J. G. & Shaw, C. G. (1955). Fungi in domestic sewage-treatment plants. *Applied microbiology* **3**, 173.
- Blinkova, O., Rosario, K., Li, L., Kapoor, A., Slikas, B., Bernardin, F., Breitbart, M. & Delwart, E. (2009). Frequent detection of highly diverse variants of cardiovirus, cosavirus, bocavirus, and circovirus in sewage samples collected in the United States. *Journal of clinical microbiology* **47**, 3507-3513.
- Blomqvist, S., Savolainen, C., Laine, P., Hirttiö, P., Lamminsalo, E., Penttilä, E., Jöks, S., Roivainen, M. & Hovi, T. (2004). Characterization of a Highly Evolved Vaccine-Derived Poliovirus Type 3 Isolated from Sewage in Estonia. *Journal of Virology* **78**, 4876-4883.
- Bodewes, R., van der Giessen, J., Haagmans, B. L., Osterhaus, A. D. M. E. & Smits, S. L. (2013). Identification of multiple novel novel viruses in feces of red foxes including a parvovirus and hepevirus. *Journal of Virology* **87**, 7758-7764.
- Boni, M. F., Posada, D. & Feldman, M. W. (2007). An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176**, 1035-1047.
- Cantalupo, P. G., Calgua, B., Zhao, G., Hundesa, A., Wier, A. D., Katz, J. P., Grabe, M., Hendrix, R. W., Girones, R., Wang, D. & Pipas, J. M. (2011). Raw Sewage Harbors Diverse Viral Populations. *mBio* **2**, e00180-00111.
- Castrignano, S. B., Nagasse-Sugahara, T. K., Kisielius, J. J., Ueda-Ito, M., Brandão, P. E. & Curti, S. P. (2013). Two novel circo-like viruses detected in human feces: complete genome sequencing and electron microscopy analysis. *Virus Research* **178**, 364-373.
- Collin, S., Fernández-lobato, M., Gooding, P. S., Mullineaux, P. M. & Fenoll, C. (1996). The two nonstructural proteins from wheat dwarf virus involved in viral gene expression and replication are retinoblastoma-binding proteins. *Virology* **219**, 324-329.
- Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164-1165.
- Dayaram, A., Galatowitsch, M., Harding, J. S., Argüello-Astorga, G. R. & Varsani, A. (2014). Novel circular DNA viruses identified in *Procordulia grayi* and *Xanthocnemis zealandica* larvae using metagenomic approaches. *Infection, Genetics and Evolution* **22**, 134-141.
- Dayaram, A., Opong, A., Jäschke, A., Hadfield, J., Baschiera, M., Dobson, R. C. J., Offei, S. K., Shepherd, D. N., Martin, D. P. & Varsani, A. (2012). Molecular

- characterisation of a novel cassava associated circular ssDNA virus. *Virus Research* **166**, 130-135.
- Dayaram, A., Potter, K. A., Moline, A. B., Rosenstein, D. D., Marinov, M., Thomas, J. E., Breitbart, M., Rosario, K., Argüello-Astorga, G. R. & Varsani, A. (2013).** High global diversity of cycloviruses amongst dragonflies. *Journal of General Virology* **94**, 1827-1840.
- Dayaram, A., Potter, K. A., Pailes, R., Moline, A. B., Marinov, M., Rosenstein, D. D. & Varsani, A. (In review).** Identification of diverse circular Rep-encoding DNA viruses in dragonflies and damselflies of Arizona and Oklahoma.
- de Jong, M. D., Van Kinh, N., Trung, N. V., Taylor, W., Wertheim, H. F., van der Ende, A., van der Hoek, L., Canuti, M., Crusat, M. & Sona, S. (2014).** Limited geographic distribution of the novel cyclovirus CyCV-VN. *Scientific reports* **4**.
- Dekker, E. L., Woolston, C. J., Xue, Y., Cox, B. & Mullineaux, P. M. (1991).** Transcript mapping reveals different expression strategies for the bicistronic RNAs of the geminivirus wheat dwarf virus. *Nucleic Acids Research* **19**, 4075-4081.
- Diemer, G. S. & Stedman, K. M. (2012).** A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. *Biol Direct* **7**, 13.
- Dorcas, M., Rani, I. S. & Sulakshan, A. (2013).** Sewage water fungi. *Ecology, Environment and Conservation* **19**, 351-352.
- Du, Z., Tang, Y., Zhang, S., She, X., Lan, G., Varsani, A. & He, Z. (2014).** Identification and molecular characterization of a single-stranded circular DNA virus with similarities to *Sclerotinia sclerotiorum* hypovirulence-associated DNA virus 1. *Arch Virol*, 1-5.
- Edgar, R. C. (2004).** MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797.
- Ge, X., Li, J., Peng, C., Wu, L., Yang, X., Wu, Y., Zhang, Y. & Shi, Z. (2011).** Genetic diversity of novel circular ssDNA viruses in bats in China. *Journal of General Virology* **92**, 2646-2653.
- Ge, X., Li, Y., Yang, X., Zhang, H., Zhou, P., Zhang, Y. & Shi, Z. (2012).** Metagenomic Analysis of Viruses from Bat Fecal Samples Reveals Many Novel Viruses in Insectivorous Bats in China. *Journal of Virology* **86**, 4620-4630.
- Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. (2000).** Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573-582.
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010).** New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**, 307-321.
- Hadfield, J., Thomas, J. E., Schwinghamer, M. W., Kraberger, S., Stainton, D., Dayaram, A., Parry, J. N., Pande, D., Martin, D. P. & Varsani, A. (2012).** Molecular characterisation of dicot-infecting mastreviruses from Australia. *Virus Research* **166**, 13-22.

- Heller-Dohmen, M., Gopfert, J., Pfannstiel, J. & Spring, O. (2011).** The nucleotide sequence and genome organization of *Plasmopara halstedii* virus. *Viol J* **8**, 123.
- Henriksson, D., Tanis, R. J., Tashian, R. E. & Nyman, P. (1981).** Amino acid sequence of the coat protein subunit in satellite tobacco necrosis virus. *Journal of molecular biology* **152**, 171-179.
- Hewitt, J., Leonard, M., Greening, G. E. & Lewis, G. D. (2011).** Influence of wastewater treatment process and the population size on human virus profiles in wastewater. *Water Research* **45**, 6267-6276.
- Ismail, M. A. & Abdel-Sater, M. A. (1994).** Mycoflora inhabiting water closet environments. *Mycoses* **37**, 53-57.
- Jeske, H., Lütgemeier, M. & Preiß, W. (2001).** DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. *The EMBO journal* **20**, 6158-6167.
- Kacprzak, M., Neczaj, E. & Okoniewska, E. (2005).** The comparative mycological analysis of wastewater and sewage sludges from selected wastewater treatment plants. *Desalination* **185**, 363-370.
- Katayama, H., Haramoto, E., Oguma, K., Yamashita, H., Tajima, A., Nakajima, H. & Ohgaki, S. (2008).** One-year monthly quantitative survey of noroviruses, enteroviruses, and adenoviruses in wastewater collected from six plants in Japan. *Water Research* **42**, 1441-1448.
- Kittigul, L., Panjangampatthana, A., Rupprom, K. & Pombubpa, K. (2014).** Genetic diversity of rotavirus strains circulating in environmental water and bivalve shellfish in Thailand. *International Journal of Environmental Research and Public Health* **11**, 1299-1311.
- Koonin, E. V. & Ilyina, T. V. (1992).** Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. *The Journal of general virology* **73** 2763-2766.
- Kraberger, S., Stainton, D., Dayaram, A., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2013).** Discovery of *Sclerotinia sclerotiorum* Hypovirulence-Associated Virus-1 in Urban River Sediments of Heathcote and Styx Rivers in Christchurch City, New Zealand. *Genome Announcements* **1**, e00559-00513.
- Kraberger, S., Thomas, J. E., Geering, A. D. W., Dayaram, A., Stainton, D., Hadfield, J., Walters, M., Parmenter, K. S., van Brunschot, S., Collings, D. A., Martin, D. P. & Varsani, A. (2012).** Australian monocot-infecting mastrevirus diversity rivals that in Africa. *Virus Research* **169**, 127-136.
- Labonté, J. M. & Suttle, C. A. (2013).** Previously unknown and highly divergent ssDNA viruses populate the oceans. *The ISME journal* **7**, 2169-2177.
- Lamberto, I., Gunst, K., Müller, H., zur Hausen, H. & de Villiers, E.-M. (2014).** Mycovirus-Like DNA Virus Sequences from Cattle Serum and Human Brain and Serum Samples from Multiple Sclerosis Patients. *Genome Announcements* **2**.
- Lefevre, P., Martin, D. P., Hoareau, M., Naze, F., Delatte, H., Thierry, M., Varsani, A., Becker, N., Reynaud, B. & Lett, J.-M. (2007).** Begomovirus ‘melting pot’ in the south-west Indian Ocean islands: molecular diversity and evolution through recombination. *Journal of General Virology* **88**, 3458-3468.

- Li, L., Kapoor, A., Slikas, B., Bamidele, O. S., Wang, C., Shaukat, S., Masroor, M. A., Wilson, M. L., Ndjanga, J.-B. N., Peeters, M., Gross-Camp, N. D., Muller, M. N., Hahn, B. H., Wolfe, N. D., Triki, H., Bartkus, J., Zaidi, S. Z. & Delwart, E. (2010a).** Multiple Diverse Circoviruses Infect Farm Animals and Are Commonly Found in Human and Chimpanzee Feces. *Journal of Virology* **84**, 1674-1682.
- Li, L., Victoria, J. G., Wang, C., Jones, M., Fellers, G. M., Kunz, T. H. & Delwart, E. (2010b).** Bat Guano Virome: Predominance of Dietary Viruses from Insects and Plants plus Novel Mammalian Viruses. *Journal of Virology* **84**, 6955-6965.
- Liu, L., Davies, J. W. & Stanley, J. (1998).** Mutational analysis of bean yellow dwarf virus, a geminivirus of the genus Mastrevirus that is adapted to dicotyledonous plants. *Journal of general virology* **79**, 2265-2274.
- Liu, L., Saunders, K., Thomas, C. L., Davies, J. W. & Stanley, J. (1999).** Bean yellow dwarf virus RepA, but not Rep, binds to maize retinoblastoma protein, and the virus tolerates mutations in the consensus binding motif. *Virology* **256**, 270-279.
- Lodder, W. J. & de Roda Husman, A. M. (2005).** Presence of Noroviruses and Other Enteric Viruses in Sewage and Surface Waters in The Netherlands. *Applied and Environmental Microbiology* **71**, 1453-1461.
- Londoño, A., Riego-Ruiz, L. & Argüello-Astorga, G. (2010).** DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. *Arch Virol* **155**, 1033-1046.
- López-Bueno, A., Tamames, J., Velázquez, D., Moya, A., Quesada, A. & Alcamí, A. (2009).** High Diversity of the Viral Community from an Antarctic Lake. *Science* **326**, 858-861.
- Martin, D. & Rybicki, E. (2000).** RDP: Detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562-563.
- Martin, D. P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P. & Varsani, A. (2011a).** Recombination in Eukaryotic Single Stranded DNA Viruses. *Viruses-Basel* **3**, 1699-1738.
- Martin, D. P., Briddon, R. W. & Varsani, A. (2011b).** Recombination patterns in dicot-infecting mastreviruses mirror those found in monocot-infecting mastreviruses. *Arch Virol* **156**, 1463-1469.
- Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D. & Lefeuvre, P. (2010).** RDP3: A flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**, 2462-2463.
- Martin, D. P., Posada, D., Crandall, K. A. & Williamson, C. (2005).** A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Research and Human Retroviruses* **21**, 98-102.
- McDaniel, L. D., Rosario, K., Breitbart, M. & Paul, J. H. (2013).** Comparative metagenomics: Natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. *Environmental microbiology* **16**, 570-585.
- Metcalf, T., Melnick, J. & Estes, M. (1995).** Environmental virology: from detection of virus in sewage and water by isolation to identification by molecular biology-a trip of over 50 years. *Annual Reviews in Microbiology* **49**, 461-487.

- Ming, H. X., Zhu, L., Feng, J. F., Yang, G. & Fan, J. F. (2014). Risk Assessment of Rotavirus Infection in Surface Seawater from Bohai Bay, China. *Human and Ecological Risk Assessment* **20**, 929-940.
- Muhire, B. M., Varsani, A. & Martin, D. P. (2014). SDT: A Virus Classification Tool Based on Pairwise Sequence Alignment and Identity Calculation. *PLoS ONE* **9**, e108277.
- Nash, T. E., Dallas, M. B., Reyes, M. I., Buhrman, G. K., Ascencio-Ibañez, J. T. & Hanley-Bowdoin, L. (2011). Functional analysis of a novel motif conserved across geminivirus Rep proteins. *Journal of Virology* **85**, 1182-1192.
- Ng, T. F. F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., Oderinde, B. S., Wommack, K. E. & Delwart, E. (2012). High Variety of Known and New RNA and DNA Viruses of Diverse Origins in Untreated Sewage. *Journal of Virology* **86**, 12161-12175.
- Ng, T. F. F., Willner, D. L., Lim, Y. W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y., Rohwer, F. & Breitbart, M. (2011). Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PLoS ONE* **6**, e20579.
- Ng, T. F. F., Zhou, Y., Chen, L.-F., Shapiro, B., Stiller, M., Heintzman, P. D., Varsani, A., Kondov, N. O., Wong, W., Deng, X., Andrews, T. D., Moorman, B. J., Meulendyk, T., MacKay, G., Gilbertson, R. & Delwart, E. (2014). Preservation of viral genomes in 700-year-old caribou feces from an subarctic ice patch. *PNAS*, doi:10.1073/pnas.1410429111.
- Notredame, C., Higgins, D. G. & Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302**, 205-217.
- Orozco, B. M. & Hanley-Bowdoin, L. (1998). Conserved Sequence and Structural Motifs Contribute to the DNA Binding and Cleavage Activities of a Geminivirus Replication Protein. *Journal of Biological Chemistry* **273**, 24448-24456.
- Oswald, W. J., Gotaas, H. B., Ludwig, H. F. & Lynch, V. (1953). Algae Symbiosis in Oxidation Ponds: III. Photosynthetic Oxygenation. *Sewage and Industrial Wastes* **25**, 692-705.
- Padidam, M., Sawyer, S. & Fauquet, C. M. (1999). Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218-225.
- Parsley, L. C., Consuegra, E. J., Thomas, S. J., Bhavsar, J., Land, A. M., Bhuiyan, N. N., Mazher, M. A., Waters, R. J., Wommack, K. E. & Harper, W. F. (2010). Census of the viral metagenome within an activated sludge microbial assemblage. *Applied and Environmental Microbiology* **76**, 2673-2677.
- Phan, T., Luchsinger, V., Avendano, L., Deng, X. & Delwart, E. (2014). Cyclovirus in nasopharyngeal aspirates of Chilean children with respiratory infections. *Journal of General Virology* **95**, 922-927.
- Phan, T. G., Kapusinszky, B., Wang, C., Rose, R. K., Lipton, H. L. & Delwart, E. L. (2011). The Fecal Viral Flora of Wild Rodents. *PLoS Pathog* **7**, e1002218.
- Phan, T. G., Vo, N. P., Boros, Á., Pankovics, P., Reuter, G., Li, O. T. W., Wang, C., Deng, X., Poon, L. L. M. & Delwart, E. (2013). The Viruses of Wild Pigeon Droppings. *PLoS ONE* **8**, e72787.

- Posada, D. & Crandall, K. A. (1998).** MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817-818.
- Reuter, G., Boros, Á., Delwart, E. & Pankovics, P. (2014).** Novel circular single-stranded DNA virus from turkey faeces. *Arch Virol*, 1-4.
- Rosario, K., Dayaram, A., Marinov, M., Ware, J., Kraberger, S., Stainton, D., Breitbart, M. & Varsani, A. (2012a).** Diverse circular single-stranded DNA viruses discovered in dragonflies (Odonata: Epiprocta). *Journal of General Virology* **93**, 2668-2681.
- Rosario, K., Duffy, S. & Breitbart, M. (2009a).** Diverse circovirus-like genome architectures revealed by environmental metagenomics. *Journal of General Virology* **90**, 2418-2424.
- Rosario, K., Duffy, S. & Breitbart, M. (2012b).** A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol* **157**, 1851-1871.
- Rosario, K., Marinov, M., Stainton, D., Kraberger, S., Wiltshire, E. J., Collings, D. A., Walters, M., Martin, D. P., Breitbart, M. & Varsani, A. (2011).** Dragonfly cyclovirus, a novel single-stranded DNA virus discovered in dragonflies (Odonata: Anisoptera). *Journal of General Virology* **92**, 1302-1308.
- Rosario, K., Nilsson, C., Lim, Y. W., Ruan, Y. & Breitbart, M. (2009b).** Metagenomic analysis of viruses in reclaimed water. *Environmental Microbiology* **11**, 2806-2820.
- Rosario, K., Symonds, E. M., Sinigalliano, C., Stewart, J. & Breitbart, M. (2009c).** Pepper Mild Mottle Virus as an Indicator of Fecal Pollution. *Applied and Environmental Microbiology* **75**, 7261-7267.
- Roux, S., Enault, F., Bronner, G., Vaultot, D., Forterre, P. & Krupovic, M. (2013).** Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. *Nature communications* **4**.
- Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime-Ngando, T. & Debroas, D. (2012).** Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. *PLoS ONE* **7**, e33641.
- Sachsenröder, J., Twardziok, S., Hammerl, J. A., Janczyk, P., Wrede, P., Hertwig, S. & Johne, R. (2012).** Simultaneous Identification of DNA and RNA Viruses Present in Pig Faeces Using Process-Controlled Deep Sequencing. *PLoS ONE* **7**, e34631.
- Schlindwein, A. D., Rigotto, C., Simões, C. M. O. & Barardi, C. R. M. (2010).** Detection of enteric viruses in sewage sludge and treated wastewater effluent, pp. 537-544.
- Sdiri-Loulizi, K., Hassine, M., Aouni, Z., Gharbi-Khelifi, H., Chouchane, S., Sakly, N., Neji-Guédiche, M., Pothier, P., Aouni, M. & Ambert-Balay, K. (2010).** Detection and molecular characterization of enteric viruses in environmental samples in Monastir, Tunisia between January 2003 and April 2007. *Journal of Applied Microbiology* **109**, 1093-1104.
- Seo, D. J., Lee, M. H., Son, N. R., Seo, S., Lee, K. B., Wang, X. & Choi, C. (2014).** Seasonal and regional prevalence of norovirus, hepatitis A virus, hepatitis E virus, and rotavirus in shellfish harvested from South Korea. *Food Control* **41**, 178-184.
- Sikorski, A., Dayaram, A. & Varsani, A. (2013a).** Identification of a Novel Circular DNA Virus in New Zealand Fur Seal (*Arctocephalus forsteri*) Fecal Matter. *Genome Announcements* **1**, e00558-00513.

- Sikorski, A., Massaro, M., Kraberger, S., Young, L. M., Smalley, D., Martin, D. P. & Varsani, A. (2013b).** Novel myco-like DNA viruses discovered in the faecal matter of various animals. *Virus Research* **177**, 209-216.
- Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M. & Birol, İ. (2009).** ABySS: A parallel assembler for short read sequence data. *Genome Research* **19**, 1117-1123.
- Sinton, L. W., Hall, C. H., Lynch, P. A. & Davies-Colley, R. J. (2002).** Sunlight Inactivation of Fecal Indicator Bacteria and Bacteriophages from Waste Stabilization Pond Effluent in Fresh and Saline Waters. *Applied and Environmental Microbiology* **68**, 1122-1131.
- Smith, J. M. (1992).** Analyzing the mosaic structure of genes. *Journal of Molecular Evolution* **34**, 126-129.
- Smits, S. L., Raj, V. S., Oduber, M. D., Schapendonk, C. M. E., Bodewes, R., Provacia, L., Stittelaar, K. J., Osterhaus, A. D. M. E. & Haagmans, B. L. (2013).** Metagenomic Analysis of the Ferret Fecal Viral Flora. *PLoS ONE* **8**, e71595.
- Smits, S. L., Zijlstra, E. E., van Hellemond, J. J., Schapendonk, C., Bodewes, R., Schürch, A. C., Haagmans, B. L. & Osterhaus, A. (2012).** Novel cyclovirus in human cerebrospinal fluid, Malawi, 2010-2011. *Emerging Infectious Diseases* **19**, 1511.
- Symonds, E. M., Griffin, D. W. & Breitbart, M. (2009).** Eukaryotic Viruses in Wastewater Samples from the United States. *Applied and Environmental Microbiology* **75**, 1402-1409.
- Tamaki, H., Zhang, R., Angly, F. E., Nakamura, S., Hong, P.-Y., Yasunaga, T., Kamagata, Y. & Liu, W.-T. (2012).** Metagenomic analysis of DNA viruses in a wastewater treatment plant in tropical climate. *Environmental Microbiology* **14**, 441-452.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. (2011).** MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**, 2713-2739.
- Tonani, K. A. A., Padula, J. A., Julião, F. C., Fregonesi, B. M., Alves, R. I. S., Sampaio, C. F., Beda, C. F., Hachich, E. M. & Segura-Muñoz, S. I. (2013).** Persistence of giardia, cryptosporidium, rotavirus, and adenovirus in treated sewage in São Paulo State, Brazil. *Journal of Parasitology* **99**, 1144-1147.
- Ulfig, K., Terakowski, M., Plaza, G. & Kosarewicz, O. (1996).** Keratinolytic fungi in sewage sludge. *Mycopathologia* **136**, 41-46.
- Vaidya, S. R., Chitambar, S. D. & Arankalle, V. A. (2002).** Polymerase chain reaction-based prevalence of hepatitis A, hepatitis E and TT viruses in sewage from an endemic area. *Journal of Hepatology* **37**, 131-136.
- van den Brand, J. M. A., van Leeuwen, M., Schapendonk, C. M., Simon, J. H., Haagmans, B. L., Osterhaus, A. D. M. E. & Smits, S. L. (2012).** Metagenomic Analysis of the Viral Flora of Pine Marten and European Badger Feces. *Journal of Virology* **86**, 2360-2365.

- van Doorn, H. R., Nghia, H. D. T., Chau, T. T. H., de Vries, M., Canuti, M., Deijs, M., Jebbink, M. F., Baker, S., Bryant, J. E. & Tham, N. T. (2013). Identification of a new cyclovirus in cerebrospinal fluid of patients with acute central nervous system infections. *MBio* **4**, e00231-00213.
- Van Heerden, J., Ehlers, M. M., Van Zyl, W. B. & Grabow, W. O. (2003). Incidence of adenoviruses in raw and treated water. *Water Research* **37**, 3704-3708.
- Varsani, A., Monjane, A. L., Donaldson, L., Oluwafemi, S., Zinga, I., Komba, E. K., Plakoutene, D., Mandakombo, N., Mboukoulida, J., Semballa, S., Briddon, R. W., Markham, P. G., Lett, J. M., Lefeuvre, P., Rybicki, E. P. & Martin, D. P. (2009a). Comparative analysis of Panicum streak virus and Maize streak virus diversity, recombination patterns and phylogeography. *Virology Journal* **6**, 194.
- Varsani, A., Shepherd, D. N., Dent, K., Monjane, A. L., Rybicki, E. P. & Martin, D. P. (2009b). A highly divergent South African geminivirus species illuminates the ancient evolutionary history of this family. *Virology Journal* **6**.
- Wright, E. A., Heckel, T., Groenendijk, J., Davies, J. W. & Boulton, M. I. (1997). Splicing features in maize streak virus virion- and complementary-sense gene expression. *The Plant Journal* **12**, 1285-1297.
- Yazdi, H., Heydarnejad, J. & Massumi, H. (2008). Genome characterization and genetic diversity of beet curly top Iran virus: a geminivirus with a novel nonanucleotide. *Virus Genes* **36**, 539-545.
- Yokoi, T., Takemoto, Y., Suzuki, M., Yamashita, S. & Hibi, T. (1999). The Nucleotide Sequence and Genome Organization of *Sclerophthora macrospora* Virus B. *Virology* **264**, 344-349.
- Yu, X., Li, B., Fu, Y., Jiang, D., Ghabrial, S. A., Li, G., Peng, Y., Xie, J., Cheng, J., Huang, J. & Yi, X. (2010). A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. *Proceedings of the National Academy of Sciences* **107**, 8387-8392.
- Zawar-Reza, P., Argüello-Astorga, G. R., Kraberger, S., Julian, L., Stainton, D., Broady, P. A. & Varsani, A. (2014). Diverse small circular single-stranded DNA viruses identified in a freshwater pond on the McMurdo Ice Shelf (Antarctica). *Infection, Genetics and Evolution* **26**, 132-138.

Chapter 9

Discussion and future directions

Contents

9.1	Background overview	305
9.2	Major findings	306
9.2.1	Summary.....	306
9.2.2	Mastrevirus diversity, host range and geographic distribution.....	307
9.2.3	Phylogeography and mastrevirus origins	309
9.2.4	Mechanisms of evolution.....	311
9.2.5	Discovery of highly diverse circular DNA viruses in New Zealand.....	312
9.3	Future directions	317
9.4	References.....	318

9.1 Background overview

Mastreviruses are found in several regions of the world including Asia, Europe and Australia, Africa and several surrounding islands. Over the last 10 years our knowledge regarding the diversity of mastreviruses has increased dramatically, with the discovery of new species on an almost annual basis. This is largely attributed to the improvement of molecular techniques and the development of new molecular tools (overview discussed in Chapter One). It is clear that mastreviruses are able to rapidly evolve through mechanisms of natural selection and recombination, facilitating the emergence of new variants within a short period of time (Harkins *et al.*, 2009a; Harkins *et al.*, 2009b; Krabberger *et al.*, 2012; Martin *et al.*, 2011b; Monjane *et al.*, 2011; van der Walt *et al.*, 2009; Varsani *et al.*, 2009; Varsani *et al.*, 2008b). MSV-A, has been well characterised due to its devastating effect on maize, one of the staple crops in Africa (Monjane *et al.*, 2011; Oluwafemi *et al.*, 2011; Owor *et al.*, 2007; Shepherd *et al.*, 2010; Varsani *et al.*, 2009). Grass-adapted strains of MSV and PanSV have also been well characterised, although to a lesser degree than MSV-A (Varsani *et al.*, 2009; Varsani *et al.*, 2008a). Among the dicot-infecting mastrevirus, CpCDV and TYDV have been the most extensively studied, however much of this work was carried out using serological detection assays and therefore little molecular information was previously available. A better understanding of mastreviruses epidemiology is fundamental for devising disease management strategies. Likewise there have been no documented studies investigating the presence of mastreviruses or similar circular DNA viruses in New Zealand.

The aim of this PhD research was to gain a wider understanding of the dynamics of mastreviruses, including the diversity, host and geographical range, mechanisms of evolution, inferences into historical movements and the possible origins of these viruses. Finally to address the question whether mastreviruses or related viruses are circulating in New Zealand, a viral metagenomics approach was used to identify circular ssDNA viruses from two sources in New Zealand.

9.2 Major findings

9.2.1 Summary

Through the course of this thesis research, 365 full mastrevirus genomes were recovered and analysed together with those available in GenBank to investigate mastrevirus dynamics. Chapter Two and Chapter Three highlight an investigation into the diversity of monocot-infecting mastrevirus with focus on those which infect predominantly wild grass species in Australia (n=41) and Africa (n=120). Among the 41 genomes recovered in Australia four new species were identified, PDSMV, DCSMV and two highly divergent mastrevirus species SSMV-1 and SSMV-2. The two divergent species (SSMV-1 and SSMV-2) are more closely related to the African monocot-infecting mastreviruses and a mastrevirus isolated from a dragonfly in Puerto Rico than to the other Australian mastreviruses. Based on current knowledge, Africa harbours significantly more mastrevirus species than Australia. The discoveries in Chapter Two however of several new species in a relatively small sample survey in Australia coupled with the knowledge that Africa has been the focus of mastrevirus research for many years (with <600 genomes recovered to date) is an indication that there may be a significantly higher level of mastreviruses diversity than currently known circulating in Australian grasses. Studies undertaken in both Chapter Two and Three reveal that certain mastreviruses have broad host ranges which encompass a large number of *Poaceae* species, in Africa it is MSV and PanSV and in Australian it is PSMV and CSMV (Fig 9.1). An extensive overview of the geographic distribution of mastreviruses in Africa shows many species have overlapping host ranges. MSV was documented on the island of Gran Canaria for the first time, extending the known geographic range of the African streak viruses north-west of the Sahara desert. Clear recombination patterns seen among the monocot-infecting mastreviruses are consistent with what has previously been documented in the African monocot-infecting mastreviruses (Shepherd *et al.*, 2008; Varsani *et al.*, 2009; Varsani *et al.*, 2008b).

Chapters Four, Five and Six collectively involve an in-depth investigation into the diversity and dynamics of the dicot-infecting mastreviruses. In these chapters 204 full dicot-infecting mastrevirus genomes were recovered from symptomatic pulse plants. The dataset generated as a result of the research undertaken in Chapter Four enabled, for the first time, a

phylogeographic analyses of these viruses. The analysis indicated that the likely origin of the most recent common ancestor of the dicot-infecting mastreviruses is possibly closer to Australia than any other regions which have been sampled. Chapter Five highlights the discovery of a novel Australian-like dicot-infecting mastrevirus from chickpea material collected in Pakistan, which based on our analysis is a putative new species, Chickpea yellow dwarf virus (CpYDV). This is the second dicot-infecting mastrevirus species to be found outside of Australia. In Chapter Six, a comprehensive survey of CpCDV in Sudan reveals a dominant strain circulating in Sudan and highlights the extensive intra-species recombination which is likely facilitating the emergence of several newly identified CpCDV strains.

Chapters Seven and Eight reveal the presence of a diverse range of circular DNA viruses in New Zealand associated with two sources: wild *Poaceae* spp and treated sewage material, many of which share similarities to geminiviruses.

9.2.2 Mastrevirus diversity, host range and geographic distribution

Increased sampling efforts have revealed that mastreviruses have a wider global distribution than previously thought and as a group show high diversity. Knowledge on the prevalence of some previously described species such as wild grass adapted MSVs and CSMVs was expanded on in Chapter Two and Chapter Three and several new species and strains identified. Within Australia four new species were identified, two of which are highly divergent, SSMV-1 and SSMV-2. Although Australia has been shown to harbour a level of diversity which rivals that found in Africa, to date no Australian monocot-infecting mastrevirus species have emerged as a agriculturally important pathogen like MSV-A, which emerged as a serious pathogen of maize in Africa ~150 years ago (Harkins *et al.*, 2009b; Monjane *et al.*, 2011). MSV-A apparently emerged following recombination events between two grass adapted MSV strains, this emergence was shown to be ~250 years after the introduction of maize into Africa (Harkins *et al.*, 2009b; Varsani *et al.*, 2008b). Knowing the short history of agriculture in Australia it is not surprising that mastreviruses have not yet emerged as major crop pathogens. Given the example of MSV-A and the rapid increase and expansion in the farming of cultivated grasses such as wheat and rice in Australia, virus ‘spill over’ is highly likely from indigenous and endemic plants. If a mastrevirus species was to

emerge as an important pathogen in Australia, having as much information at hand regarding these viruses is vital for surveillance and management of crop plants.

The number of dicot-infecting mastrevirus species is less than those identified to infect monocot plants. Among those identified, the majority of the dicot-infecting mastrevirus have been found only in Australia which is surprising considering research has predominantly focused on pulses in the northern hemisphere rather than in Australia. The discovery of a new species of dicot-infecting mastrevirus in Pakistan (Chapter Five), however, is the first indication that a larger pool of diversity other than those viruses belonging to the species CpCDV may exist outside of Australia. The studies undertaken in Chapter Four and Six highlight that even among CpCDV isolates analysed there is a high level of diversity, for example within Sudan alone 11 of the 14 strains were identified. This level of strain based diversity is similar to that identified for MSV and PanSV discussed in Chapter Two. It is apparent from the findings in this thesis that we have likely only just begun to unravel the true breadth of diversity within the mastrevirus genus. The discovery of several new species in Australia from a small sample suggests that further research activities in this area may identify Australia as a mastrevirus diversity hotspot.

The known host range of monocot-infecting mastreviruses includes a large number of grass species and the overall host range is much more extensive than that seen for the dicot-infecting mastreviruses. Among the dicot-infecting mastrevirus species CpCDV and TYDV have the broadest known host ranges, whereas all other species have only been identified in a single host. This is likely due to the limited amount of work which has been undertaken on wild pulses in relation to the dicot-infecting mastreviruses. An overview of monocot-infecting mastrevirus host range is shown in Figure 9.1, which illustrates the relationships between monocot-infecting mastrevirus species/strains and known host genera. Species vary in host specificity with some species and strains are more generalist having broad host ranges whereas others are more specialist with seemingly narrow host ranges. It does however need to be noted that host range can be biased by the sampling approach and/or specificity of the virus isolation method and therefore it may be that some viral species with seemly narrow host ranges do in fact have broader host ranges which is yet to be elucidated. Both MSV and PanSV have extensive host ranges which may be a reflection of the extent to which these viruses have been studied in Africa compared with other species. In Australia CSMV and

PSMV seem to be the most generalist species of monocot-infecting mastrevirus. Interestingly, several of the monocot-infecting mastrevirus species also have overlapping grass hosts. Among the species which have overlapping host ranges are also those from two geographically distinct regions. For example *Digitaria* sp. has been identified as a host for MSV, PanSV from Africa, and CSMV, DDSMV, DCSMV and PSMV from Australia. An overlap in host range can result in recombination as two viruses must infect the same host for a recombination event to take place. Therefore identifying host species that can harbour two or more viral species may give important insights into those species which are more likely to be in an environment which facilitates recombination. Having a broad knowledge of host range is also important for the management of viral infections of crops as alternative hosts can act as viral reservoirs.

Geographically mastrevirus have been found on the continents of Africa, Australia, Asia and Europe, with Africa harbouring the largest number of known monocot-infecting mastrevirus species and Australia the largest number of dicot-infecting mastrevirus species. Chapter Three highlights that within Africa some species such as MSV and PanSV are found to have wide, overlapping geographic ranges which include some of the islands off the African continent. In Australia all the identified monocot-infecting mastrevirus species have been found within the Australian state of Queensland because sampling has been so far restricted to mainly this region, it is likely that the geographical ranges of these viruses extends to other regions in Australia.

CpCDV has been identified in 11 countries whereas the other dicot-infecting mastrevirus species have only been identified in single countries. CpCDV likely has been moved between countries within the Middle East, Africa and Indian Subcontinent through the movement of infected pulse plant material and/or vector migration.

9.2.3 Phylogeography and mastrevirus origins

In Chapter Four a phylogeographic analyses was possible due to the sample size and time span over which the samples were collected. Results of this analysis indicated that the most recent common ancestor originated in the region around or possibly in Australia with the

plausible routes of movement out of Australia including two initial movements, one to Southern Africa and one to the horn of Africa. This was followed by the subsequent dispersal of these viruses to the Middle East, and Asia. The discovery of an Australian-like dicot-infecting mastrevirus in Pakistan (Chapter Five) questions the origin of this virus (CpYDV) and whether it was introduced by the movement of infected plant material by humans recently. Several other divergent mastreviruses which are more similar to those found in other geographically distant locations, such as DSV from Vanuatu which is most closely related to African streak viruses, DfasMV from Puerto Rico is most closely related to SSMV-1 and SSMV-2 from Australia. The human mediated movement of plant viruses into new regions and countries has been recently highlighted (De Bruyn *et al.*, 2012; Lefeuvre *et al.*, 2010; Ochola *et al.*, 2015; Péréfarres *et al.*, 2012). Novel mastreviruses have also been identified in quarantine sugarcane samples in France within sugarcane setts from the USA, Barbados, Sudan and Egypt. Interestingly, the sugarcane setts from the USA and Barbados all originated in Sudan (Candresse *et al.*, 2014). This information collectively highlights the possibility that DSV, DfasMV, SSMV-1 and SSMV-2 viruses may have been introduced from other regions through human mediated movement of infected plant material or insect vectors. With human movement globally being at an all-time high and trading of agricultural commodities internationally occurring more frequently, biosecurity measures to reduce the spread of plant viral pathogens between countries, are essential.

9.2.4 Mechanisms of evolution

Common patterns of recombination have often been noted among geminiviruses (Lefeuve *et al.*, 2007; Martin *et al.*, 2011a; Martin *et al.*, 2011b; Owor *et al.*, 2007; Padidam *et al.*, 1999; Silva *et al.*, 2014; van der Walt *et al.*, 2009; Varsani *et al.*, 2009; Varsani *et al.*, 2008b). It is evident from the recombination analyses undertaken as part of the work presented in this thesis that recombination is a major driving force behind the diversification of mastreviruses and facilitating the emergence of new strains / variants and possibly new species. Clear patterns seen among mastreviruses in the analysis shown in Chapters Two, Three, Four and Six are:

- 1) The size of genetic fragment exchanged in intra-species recombination tends to be larger on average than that seen in inter-species, with the exception of analyses undertaken in Chapter Six where fragments were considerably larger than described in other analyses (Chapters Two, Three and Four).
- 2) Clear recombination breakpoint hotspots are noted in the intergenic regions and more frequently in the complementary-sense genes than in the virion-sense genes.
- 3) The first evidence for inter-species recombination events between species from two geographically distant locations was identified in the monocot-infecting mastreviruses (Chapter Three) and the dicot-infecting mastreviruses (Chapters Four and Six).
- 4) Evidence that some of these species at one time must have circulated in the same geographic location and occupied the same host(s).

An investigation into the selection pressures acting on the coding regions of monocot and dicot-infecting mastrevirus species showed that the same genes are evolving under similar selection pressures between species. Overall, purifying selection is dominant implying these genes are predominantly favouring the maintenance of many codon sites. A region which generally is evolving under largely purifying in the *cp* of all mastreviruses analysed is within the possible β -barrel motif region which forms the core structure of the *cp*. No regions are evidently evolving under similar selection pressures across all species of dicot- and monocot-infecting mastreviruses. There are however some selection signals trends between species within the two groupings (monocot and dicot-infecting) that are highlighted in Chapter Three

and Six. Interestingly, certain sites among species which have members that are known to infect a wide range of hosts such as MSV and PanSV have sites which are undergoing episodic selection at codons clustered within certain regions of the *cp* and *rep*, which could be an indication that these sites may play a role in host specificity. This may be useful for future studies which are interested in target regions for investigating host and vector specificity.

9.2.5 Discovery of highly diverse circular DNA viruses in New Zealand

Several studies have used sequence-independent approaches such as viral metagenomics to identify viral populations in a sample without any prior knowledge of what viruses are present (Rosario *et al.*, 2012). Many of which have led to the discovery of novel geminiviruses and gemini-like viruses (Candresse *et al.*, 2014; Du *et al.*, 2014; Kraberger *et al.*, 2013b; Kreuze *et al.*, 2009; Loconsole *et al.*, 2012; Ng *et al.*, 2011a; Ng *et al.*, 2012; Ng *et al.*, 2011b; Poojari *et al.*, 2013; Seguin *et al.*, 2014; Sikorski *et al.*, 2013). A recent publication has also demonstrated the effectiveness of viral metagenomics in the detection of novel mastrevirus in quarantine samples (Candresse *et al.*, 2014). Plant viruses such as pepper mild mottle virus have been shown to be viable in sewage treatment processes and can act as an indicator of faecal contaminated water (Rosario *et al.*, 2009).

In light of these discoveries an exploratory approach undertaken in Chapter Seven and Eight used a viral metagenomic approach to identify novel mastreviruses or similar CRESS DNA viruses in two sources, wild grasses and a sewage oxidation pond. A sample from an oxidation pond was chosen instead of untreated sewage because we do not have access to facilities that allows us to work on the latter and it was of particular interest to investigate whether such viruses were present following the treatment process prior to discharge back into the local environment. Both studies employed a viral metagenomics approach using NGS coupled with the use of back-to-back primers for PCR amplification and recovery of full clonal viral genomes. Viral genomic DNA was enriched using Phi29 polymerase prior to sequencing which enhances the discovery of circular DNA molecules. Although no mastreviruses were recovered, four novel circular ssDNA virus genomes were isolated from grasses sampled on both the North and South Islands of New Zealand and 50 novel CRESS

DNA viruses from a sewage oxidation pond sample, of which many share similarities to geminiviruses.

The four novel CRESS DNA viruses associated with grasses in New Zealand. BasCV-1 and BasCV-3 have likely replication-associated proteins which are expressed following a splicing event and similar genome architecture to that of mastreviruses, although a possible movement protein was not identified. These novel viruses represent new highly divergent species and possibly belong to new genera. The putative Rep of BasCV-1 shares between 29-32% identity with the Rep of the Nepavirus from raw sewage, SaCV-4 from treated sewage and other geminiviruses. The Rep of BasCV-2 shares 31-34% identity to that of Rodent stool associated circular virus M-45 and environmentally derived ssDNA viruses. BaCV-3 phylogenetically clusters with the Gemycircularviruses and interestingly is most closely related to two other genomes recovered from leaf material and one from the fungus *Sclerotinia sclerotiorum*. All BasCVs viruses were identified associated with *Bromus* grasses and BasCV-1 was found in two locations within New Zealand. This information indicates these likely have are widely distributed within New Zealand and lends weight to the association of this virus with *Bromus* spp in New Zealand. It is possible that these viruses infect grasses or an organism such as fungi, bacteria or protists which are associated with the grass.

Following on from Chapter Seven the body of work in Chapter Eight uses a similar approach to identify CRESS DNA viruses in a sewage oxidation pond sample from Christchurch, New Zealand. A range of highly diverse CRESS DNA viruses were recovered, many of which are distantly related to geminiviruses (SaCV-1 – SaCV-4, SaCV-36 and SaCV-37) as well as other major CRESS DNA virus families including circoviruses, cycloviruses and nanoviruses. Thirteen genomes discovered cluster with the gemycircularviruses, expanding this group and enabling a more in-depth analysis. The gemycircularvirus genus has been populated with members recovered from a wide range of sources such as fungi, insects, plants, faecal material, bovine serum and human brain tissue, leading to the question of whether these viruses are able to move between organisms in different kingdoms. The recent discovery of a common algal virus, ATCV-1, associated with human mycosal surfaces is an

example of how viruses can be part of the virome of organisms that are not within the same kingdom as the known host (Yolken *et al.*, 2014).

Several new gemycircularviruses were identified collectively in Chapter Seven and Eight, this meant that for the first time a rigorous recombination analysis could be undertaken. It is evident from this analysis that recombination is likely a frequent occurrence, as is seen among mastreviruses (Chapter Two, Three, Four and Six) and other ssDNA viruses (Hadfield *et al.*, 2012; Krabberger *et al.*, 2013a; Martin *et al.*, 2011a; Stainton *et al.*, 2012; Stenzel *et al.*, 2014). The gemycircularvirus dataset however is still relatively small and therefore no obvious trends in recombination patterns or hot/cold spots could be identified.

Putative nonanucleotide, RCR and SF3 motifs were identified for all the novel viral genomes identified in Chapter Seven and Eight. The Reps of several of these viral genomes identified are related those of geminiviruses. These genomes also each contain a nonanucleotide sequence which are similar or homologous to those of geminiviruses. For example those viruses belonging to the gemycircularvirus group all contained a well conserved nonanucleotide motif of TAATR TTAD. There is also a high degree of conservation of the RCR and SF3 motifs between these gemini-like viruses and geminiviruses. This is also the case for other grouping/clades where evidence of motif conservation is beginning to become obvious, such as that seen in CRESS DNA virus clade-1 (Chapter Eight). These similarities and phylogenetic relationships to other known ssDNA virus families such as geminiviruses could be clue to the evolutionary origins of these novel viruses. As more information becomes available on these viruses and phylogenetic groups are expanded we may be able to elucidate the origins of all ssDNA viruses.

Infection studies to identify potential hosts of these viruses recovered in Chapters Seven and Eight could not to be performed as part of this thesis research due to time constraints, permits and availability of appropriate containment facilities. Nonetheless the discoveries of more than 50 novel CRESS DNA viruses in Chapters Seven and Eight, many of which share similarities to geminiviruses, further iterates that such sequence independent approaches are ideal for the discovery of novel gemini-like viruses and provides insights into the wealth of

CRESS DNA diversity present in different environments within New Zealand. A wide range of highly diverse CRESS DNA viruses were identified from a relatively small dataset of 33 grass samples and a single sample from an oxidation pond, therefore it is likely that expanding this approach to look at other sources would result in the discovery of many more gemini-like viruses and novel CRESS DNA viruses.

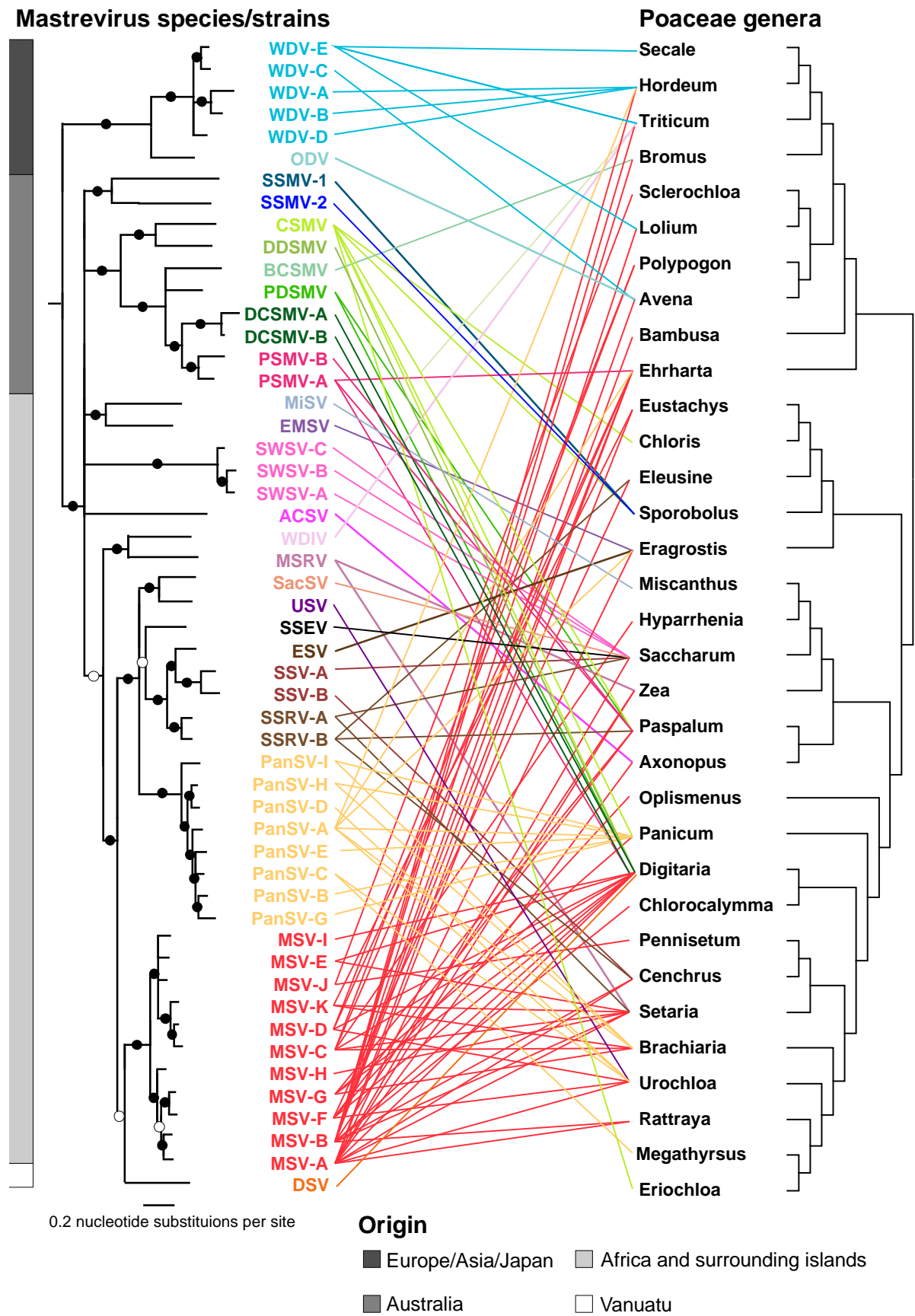


Figure 9.1: Depiction of monocot-infecting mastrevirus species and strains and the corresponding host genera from which full genomes have been recovered.

9.3 Future directions

As part of the research undertaken in this thesis we have expanded the current knowledge of mastreviruses globally and novel CRESS DNA viruses in New Zealand. This has opened the field up to new questions and has highlighted areas where more research is necessary. These areas include the following:

- Transmission studies to identify the leafhopper species transmitting the Australian dicot-infecting mastreviruses, CpYDV in Pakistan and the monocot-infecting mastreviruses around the world.
- Survey of geographical distribution of leafhopper species in pulse growing regions of Australia, Asia and Africa to gain a better perspective of vector dynamics. In order to identify vector species visual identification coupled with molecular analysis of a conserved gene such as the *cytochrome c oxidase I* gene could be used. This could also solve the debate as to whether *O. orientalis* and *O. albicinctus* are different species or the same.
- Use CpYDV specific primers to investigate the presence of this species in major pulse growing regions of the world and to possibly answer the question to its likely origin.
- Identify dicot-infecting mastreviruses in non-cultivated plants and determine whether these act as reservoirs and potential ‘mixing vessels’ for the emergence of recombinant variants that could pose serious threat to cultivated crops.
- Undertake an extensive survey of indigenous grasses in Australia to illuminate the diversity of monocot-infecting mastreviruses in Australia.
- Investigate potential hosts of the novel CRESS DNA viruses recovered in this study by designing specific probes, performing infection studies and using a sequence based approach.

9.4 References

- Candresse, T., Filloux, D., Muhire, B., Julian, C., Galzi, S., Fort, G., Bernardo, P., Daugrois, J.-H., Fernandez, E., Martin, D. P., Varsani, A. & Roumagnac, P. (2014). Appearances Can Be Deceptive: Revealing a Hidden Viral Infection with Deep Sequencing in a Plant Quarantine Context. *PLoS ONE* **9**, e102945.
- De Bruyn, A., Villemot, J., Lefeuvre, P., Villar, E., Hoareau, M., Harimalala, M., Abdoul-Karime, A. L., Abdou-Chakour, C., Reynaud, B. & Harkins, G. W. (2012). East African cassava mosaic-like viruses from Africa to Indian ocean islands: molecular diversity, evolutionary history and geographical dissemination of a bipartite begomovirus. *BMC evolutionary biology* **12**, 228.
- Du, Z., Tang, Y., Zhang, S., She, X., Lan, G., Varsani, A. & He, Z. (2014). Identification and molecular characterization of a single-stranded circular DNA virus with similarities to *Sclerotinia sclerotiorum* hypovirulence-associated DNA virus 1. *Arch Virol* **159**, 1527-1531.
- Hadfield, J., Thomas, J. E., Schwinghamer, M. W., Krabberger, S., Stainton, D., Dayaram, A., Parry, J. N., Pande, D., Martin, D. P. & Varsani, A. (2012). Molecular characterisation of dicot-infecting mastreviruses from Australia. *Virus Research* **166**, 13-22.
- Harkins, G. W., Delport, W., Duffy, S., Wood, N., Monjane, A. L., Owor, B. E., Donaldson, L., Saumtally, S., Triton, G., Briddon, R. W., Shepherd, D. N., Rybicki, E. P., Martin, D. P. & Varsani, A. (2009a). Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts. *Virology Journal* **6**.
- Harkins, G. W., Martin, D. P., Duffy, S., Monjane, A. L., Shepherd, D. N., Windram, O. P., Owor, B. E., Donaldson, L., van Antwerpen, T., Sayed, R. A., Flett, B., Ramusi, M., Rybicki, E. P., Peterschmitt, M. & Varsani, A. (2009b). Dating the origins of the maize-adapted strain of maize streak virus, MSV-A. *Journal of General Virology* **90**, 3066-3074.
- Krabberger, S., Harkins, G. W., Kumari, S. G., Thomas, J. E., Schwinghamer, M. W., Sharman, M., Collings, D. A., Briddon, R. W., Martin, D. P. & Varsani, A. (2013a). Evidence that dicot-infecting mastreviruses are particularly prone to inter-species recombination and have likely been circulating in Australia for longer than in Africa and the Middle East. *Virology* **444**, 282-291.
- Krabberger, S., Stainton, D., Dayaram, A., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2013b). Discovery of *Sclerotinia sclerotiorum* Hypovirulence-Associated Virus-1 in Urban River Sediments of Heathcote and Styx Rivers in Christchurch City, New Zealand. *Genome Announcements* **1**, e00559-00513.
- Krabberger, S., Thomas, J. E., Geering, A. D. W., Dayaram, A., Stainton, D., Hadfield, J., Walters, M., Parmenter, K. S., van Brunschot, S., Collings, D. A., Martin, D. P. & Varsani, A. (2012). Australian monocot-infecting mastrevirus diversity rivals that in Africa. *Virus Research* **169**, 127-136.
- Kreuze, J. F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I. & Simon, R. (2009). Complete viral genome sequence and discovery of novel viruses by deep

- sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. *Virology* **388**, 1-7.
- Lefevre, P., Martin, D. P., Harkins, G., Lemey, P., Gray, A. J., Meredith, S., Lakay, F., Monjane, A., Lett, J.-M. & Varsani, A. (2010).** The spread of Tomato yellow leaf curl virus from the Middle East to the world. *PLoS Pathogens* **6**, e1001164.
- Lefevre, P., Martin, D. P., Hoareau, M., Naze, F., Delatte, H., Thierry, M., Varsani, A., Becker, N., Reynaud, B. & Lett, J.-M. (2007).** Begomovirus ‘melting pot’ in the south-west Indian Ocean islands: molecular diversity and evolution through recombination. *Journal of General Virology* **88**, 3458-3468.
- Loconsole, G., Saldarelli, P., Doddapaneni, H., Savino, V., Martelli, G. P. & Saponari, M. (2012).** Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease, a new member in the family Geminiviridae. *Virology* **432**, 162-172.
- Martin, D. P., Biagini, P., Lefevre, P., Golden, M., Roumagnac, P. & Varsani, A. (2011a).** Recombination in Eukaryotic Single Stranded DNA Viruses. *Viruses* **3**, 1699-1738.
- Martin, D. P., Briddon, R. W. & Varsani, A. (2011b).** Recombination patterns in dicot-infecting mastreviruses mirror those found in monocot-infecting mastreviruses. *Arch Virol* **156**, 1463-1469.
- Monjane, A. L., Harkins, G. W., Martin, D. P., Lemey, P., Lefevre, P., Shepherd, D. N., Oluwafemi, S., Simuyandi, M., Zinga, I., Komba, E. K., Lakoutene, D. P., Mandakombo, N., Mboukoulida, J., Semballa, S., Tagne, A., Tiendrebeogo, F., Erdmann, J. B., van Antwerpen, T., Owor, B. E., Flett, B., Ramusi, M., Windram, O. P., Syed, R., Lett, J. M., Briddon, R. W., Markham, P. G., Rybicki, E. P. & Varsani, A. (2011).** Reconstructing the history of maize streak virus strain a dispersal to reveal diversification hot spots and its origin in southern Africa. *Journal of Virology* **85**, 9623-9636.
- Ng, T. F. F., Duffy, S., Polston, J. E., Bixby, E., Vallad, G. E. & Breitbart, M. (2011a).** Exploring the Diversity of Plant DNA Viruses and Their Satellites Using Vector-Enabled Metagenomics on Whiteflies. *PLoS ONE* **6**, e19050.
- Ng, T. F. F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., Oderinde, B. S., Wommack, K. E. & Delwart, E. (2012).** High Variety of Known and New RNA and DNA Viruses of Diverse Origins in Untreated Sewage. *Journal of Virology* **86**, 12161-12175.
- Ng, T. F. F., Willner, D. L., Lim, Y. W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y., Rohwer, F. & Breitbart, M. (2011b).** Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PLoS ONE* **6**, e20579.
- Ochola, D., Issaka, S., Rakotomalala, M., Pinel-Galzi, A., Ndikumana, I., Hubert, J., Hébrard, E., Séré, Y., Tusiime, G. & Fargette, D. (2015).** Emergence of rice yellow mottle virus in eastern Uganda: Recent and singular interplay between strains in East Africa and in Madagascar. *Virus Research* **195**, 64–72.
- Oluwafemi, S., Alegbejo, M. D., Onasanya, A. & Olufemi, O. (2011).** Relatedness of Maize streak virus in maize (*Zea mays* L.) to some grass isolates collected from different regions in Nigeria. *African Journal of Agricultural Research* **6**, 5878-5883.
- Owor, B. E., Martin, D. P., Shepherd, D. N., Edema, R., Monjane, A. L., Rybicki, E. P., Thomson, J. A. & Varsani, A. (2007).** Genetic analysis of maize streak virus isolates

- from Uganda reveals widespread distribution of a recombinant variant. *Journal of General Virology* **88**, 3154-3165.
- Padidam, M., Sawyer, S. & Fauquet, C. M. (1999).** Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218-225.
- Péréfarres, F., De Bruyn, A., Krabberger, S., Hoareau, M., Barjon, F., Lefeuvre, P., Pellegrin, F., Caplong, P., Varsani, A. & Lett, J. (2012).** Occurrence of the Israel strain of Tomato yellow leaf curl virus in New Caledonia and Loyalty Islands. *New Disease Reports* **25**.
- Poojari, S., Alabi, O. J., Fofanov, V. Y. & Naidu, R. A. (2013).** A leafhopper-transmissible DNA virus with novel evolutionary lineage in the family Geminiviridae implicated in grapevine redleaf disease by next-Generation sequencing. *PloS one* **8**, e64194.
- Rosario, K., Duffy, S. & Breitbart, M. (2012).** A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol* **157**, 1851-1871.
- Rosario, K., Symonds, E. M., Sinigalliano, C., Stewart, J. & Breitbart, M. (2009).** Pepper Mild Mottle Virus as an Indicator of Fecal Pollution. *Applied and Environmental Microbiology* **75**, 7261-7267.
- Seguin, J., Rajeswaran, R., Malpica-López, N., Martin, R. R., Kasschau, K., Dolja, V. V., Otten, P., Farinelli, L. & Pooggin, M. M. (2014).** *De Novo* Reconstruction of Consensus Master Genomes of Plant RNA and DNA Viruses from siRNAs. *PLoS ONE* **9**, e88513.
- Shepherd, D. N., Martin, D. P., Van Der Walt, E., Dent, K., Varsani, A. & Rybicki, E. P. (2010).** Maize streak virus: An old and complex 'emerging' pathogen. *Molecular Plant Pathology* **11**, 1-12.
- Shepherd, D. N., Varsani, A., Windram, O. P., Lefeuvre, P., Monjane, A. L., Owor, B. E. & Martin, D. P. (2008).** Novel sugarcane streak and sugarcane streak Reunion mastreviruses from southern Africa and la Réunion. *Arch Virol* **153**, 605-609.
- Sikorski, A., Massaro, M., Krabberger, S., Young, L. M., Smalley, D., Martin, D. P. & Varsani, A. (2013).** Novel myco-like DNA viruses discovered in the faecal matter of various animals. *Virus Research* **177**, 209-216.
- Silva, F. N., Lima, A. T., Rocha, C. S., Castillo-Urquiza, G. P., Alves-Júnior, M. & Zerbini, F. M. (2014).** Recombination and pseudorecombination driving the evolution of the begomoviruses Tomato severe rugose virus (ToSRV) and Tomato rugose mosaic virus (ToRMV): two recombinant DNA-A components sharing the same DNA-B. *Virology journal* **11**, 66.
- Stainton, D., Krabberger, S., Walters, M., Wiltshire, E. J., Rosario, K., Halafihi, M., Lolohea, S., Katoa, I., Faitua, T. H., Aholelei, W., Taufu, L., Thomas, J. E., Collings, D. A., Martin, D. P. & Varsani, A. (2012).** Evidence of inter-component recombination, intra-component recombination and reassortment in banana bunchy top virus. *The Journal of general virology* **93**, 1103-1119.
- Stenzel, T., Piasecki, T., Chrzastek, K., Julian, L., Muhire, B. M., Golden, M., Martin, D. P. & Varsani, A. (2014).** Pigeon circoviruses display patterns of recombination, genomic secondary structure and selection similar to those of beak and feather disease viruses. *Journal of General Virology* **95**, 1338-1351.
- van der Walt, E., Rybicki, E. P., Varsani, A., Polston, J. E., Billharz, R., Donaldson, L., Monjane, A. L. & Martin, D. P. (2009).** Rapid host adaptation by extensive recombination. *Journal of General Virology* **90**, 734-746.

- Varsani, A., Monjane, A. L., Donaldson, L., Oluwafemi, S., Zinga, I., Komba, E. K., Plakoutene, D., Mandakombo, N., Mboukoulida, J., Semballa, S., Briddon, R. W., Markham, P. G., Lett, J. M., Lefeuvre, P., Rybicki, E. P. & Martin, D. P. (2009).** Comparative analysis of Panicum streak virus and Maize streak virus diversity, recombination patterns and phylogeography. *Virology Journal* **6**, 194.
- Varsani, A., Oluwafemi, S., Windram, O., Shepherd, D., Monjane, A., Owor, B., Rybicki, E., Lefeuvre, P. & Martin, D. (2008a).** Panicum streak virus diversity is similar to that observed for maize streak virus. *Arch Virol* **153**, 601-604.
- Varsani, A., Shepherd, D. N., Monjane, A. L., Owor, B. E., Erdmann, J. B., Rybicki, E. P., Peterschmitt, M., Briddon, R. W., Markham, P. G., Oluwafemi, S., Windram, O. P., Lefeuvre, P., Lett, J. M. & Martin, D. P. (2008b).** Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. *Journal of General Virology* **89**, 2063-2074.
- Volken, R. H., Jones-Brando, L., Dunigan, D. D., Kannan, G., Dickerson, F., Severance, E., Sabunciyan, S., Talbot, C. C., Prandovszky, E., Gurnon, J. R., Agarkova, I. V., Leister, F., Gressitt, K. L., Chen, O., Deuber, B., Ma, F., Pletnikov, M. V. & Van Etten, J. L. (2014).** Chlorovirus ATCV-1 is part of the human oropharyngeal virome and is associated with changes in cognitive functions in humans and mice. *Proceedings of the National Academy of Sciences* **111**, 16106-16111.